Bounded Rationality in Communication

Jan Willem Wennekes
(1013823)
July 2, 2002

Thesis Advisors
Dr. Petra Hendriks (RuG)
Dr. Niels Taatgen (RuG)
Dr. Rineke Verbrugge (RuG)

Artificial Intelligence
University of Groningen

Bounded Rationality in Communication

Student
Jan Willem Wennekes – 1013823
stinger@funsport.nl

Thesis advisors
Dr. Petra Hendriks (petra@ai.rug.nl)
Dr. Rineke Verbrugge (rineke@ai.rug.nl)
Dr. Niels Taatgen (niels@ai.rug.nl)
Artificial Intelligence
University of Groningen

July 2, 2002

Artificial Intelligence
University of Groningen

Contents

1. Introduction

The history of the study of Intelligence can be traced back to the ancient times. Humans have always been anxious to find out how the human mind works. The ancient Greeks developed a system of reasoning (rhetorics, logics and dialectics) to describe the process of normal human reasoning. The development of ancient practices of philosophy into different autonomous sciences such as philosophy, mathematics, psychology and linguistics has fed the interest in the human mind as well as our knowledge about our brains and the processes within them.

During this development of the studies of intelligence, the relatively recent development of computers brought new and interesting ideas. One of the first ideas when the first computers came around, was the comparison of the logical circuitry of computers with the workings of interconnected neurons in the brain. This idea from McCulloch and Pitts (McCulloch & Pitts, 1943) was very influential because of the analogy between a computer system and the human brain. They claimed that the logical properties of the brain as a whole could be understood in terms of the logical properties of its constituent cells. A new field of research and development was born; Artificial Intelligence (AI) aimed at creating working models of human intelligence, using theories and methods from different studies of intelligence (psychology, neurology, philosophy, mathematics, linguistics, etc).

This view on cognition is what I call a computational vision. The mind is a very complex system, which is able to compute many different problems in different areas of our lives. In the computational view, the workings of the human mind are seen as computation. The mind is in fact a very complex 'calculator'. Most AI models are based on this computational vision; the mind has perfect information, limitless computational power and lots of time to make decisions. This can be seen in many examples of AI models and programs, e.g. chess programs. The great power behind these programs comes not from the truly intelligent model, but from the great computing speed and ease of modern computer systems.

Even though these computational models have supplied interesting and good programs, the mind is still erroneously being modeled as being a computer. In fact, these models can be seen as prescriptions of the cognitive processes at investigation rather than descriptions of the actual working of the mind. Even the formal systems of reasoning and logic that have developed since the ancient Greeks seem to be abstractions of normal human cognition. Von Neumann, who stood at the beginning of modern AI, doubted if logic and mathematics could eventually model human thought: "the language of the brain is not the language of mathematics", as he put it (Von Neumann, 1958, p, 80).

It appears that humans make inferences about the world in a different way. We have to deal with limited knowledge, time and computational power. Our cognitive capabilities do not include enormous computational and logical powers. This aspect of real-life cognition has been described as bounded rationality. We make inferences about the world in a limited timespan, based upon little information. Because of our limited capacities, we have developed certain strategies that make optimal use of the resources available. This can be seen as an ecological rationality (Todd & Gigerenzer, 2000). This ecological rationality consists of fast and frugal heuristics that have evolved in our minds and societies up till today. These heuristics have to be learned and result from the fit between the mind's mechanisms and the structure of the environment in which it operates.

Bounded rationality can be found in many areas of cognition in everyday life: a doctor who decides which treatment to give a patient in the emergency room in a hospital; a surfer who decides which sail to use today, etc. One area where we expect to find examples of bounded rationality is the domain of language. In communication people have to make fast decisions about the meaning and use of words.

In this project I will look at the generation and interpretation of nominal referring expressions, such as *the circle, a blue circle*. Such expressions are widely used in discourse in everyday life. Different nominal referring expressions can be used to refer to the same object, but the actual choice depends on various factors. How do humans determine which possible nominal referring expression is the best to be used? And, on the other hand, how do humans know which object or entity is meant by the used nominal referring expression?

The key objective is to test whether humans are restricted or aided by their bounded rationality in finding an optimal strategy. I will use a working cognitive model capable of communicating with a human subject in a graph completion task. In this task participants have to complete a graph, consisting of several colored circles. Participants have to cooperate with an ACT-R model in order to complete the graph, since the individual participants only have part of the graph.

The organization of this work is as follows:

Theoretical Background
Here I will lay down a theoretical background against which we can later on discuss my goals, methods and results.

Scientific Goals
This chapter will describe what I want to achieve with this project. I will state my research question here and present my hypothesis. I will also describe the scientific relevance of this project for Artificial Intelligence.

Methods

I will give a detailed description of the model I used and of the experimental setting in this chapter. I will also give predictions for the outcome of the experiment, based on my hypothesis and the theoretical background.

Results

This chapter will reflect the results from the experiments. I will present the data here in the form of graphs with explanatory comments.

Discussion

Here I will give an interpretation of the data. Are the effects as I expected? And what does this mean for my hypothesis? I will also give an evaluation and discuss some of the participants' suggestions for the future. Since this project is a pilot-study, I will give some suggestions for further research.

Conclusions

In this chapter I will conclude with final remarks about my research and give a summary of what I achieved.

2. Theoretical Background

2.1. Introduction

In this chapter I will discuss the different theories I will use for my research. First I will discuss bounded rationality. This theory incorporates the limitations of the human cognitive system. Next I will discuss Optimality Theory, which is a formal theory of the workings of natural language. Furthermore I discuss several theoretical subjects that I need for my research and experiments, e.g. referring expressions and common ground.

2.2. Bounded Rationality

2.2.1. Decision making

In everyday life we have to make decisions. Some are important and can affect our lives greatly, while others are arbitrary. It seems quite normal for us to make decisions, and we can make a lot of decisions very fast and effectively. When, e.g., driving in a car, we can make decisions about steering, speeding and braking in an instant, effectively taking care of our security and the security of others. In different areas of our lives we use our decision-making capabilities daily. For important and influential choices, but also for simple choices such as what to eat for lunch today.

In theories of rational decision making, such as rational choice theory, humans are regarded as efficient computational machines that maximize a measure of expected utility that reflects a complete and consistent preference order and probability measure over all possible choices (Doyle, 1999). This is the way in which many artificial intelligence programs, such as chess-programs, model decision making. Vast processes of computation explain our fast decisions. But everyday decision making does not live up to this high rational standard. Even expert decision-makers do not make decisions in the spirit of rational choice theory (e.g. Kahneman, Slovic & Tversky, 1982; Machina, 1987). In real life, things seem to be different. The complexity of decision making does not seem to be such a vast process of computation. Humans cannot compute all possible options and pay-offs, because of several factors. First, in most everyday situations there is risk and uncertainty. Second, humans normally do not have complete information about the situation and the alternatives. Third, the situation or environment might be too complex to calculate the best course of action. These three factors complicate the process of computing the optimal choice (Simon, 1972). In effect, these factors make decisions so complicated – if one wants to take all factors into account - that it would take humans an extensive, impractical amount of time to make even a simple decision.

2.2.2. Bounded rationality

Rather than unbounded optimization, people seem to be best described by theories of bounded rationality (Simon, 1972; 1983; 1997; Todd & Gigerenzer, 1999; 2000). These theories do not describe humans as computational machines that try to optimize their utility, but take into account the constraints on the human cognitive system as noted above. The following three aspects of decision-making seem to be incompatible with theories of rational decision making (Simon 1983). I will explain these three aspects in the not so complex context of choosing your lunch for today.

- F1 - First, decisions are not about your life as a whole, but tend to concern specific aspects of your life that are relatively independent of each other. When, e.g., you choose what to have for lunch today, you do not decide what you are going to wear to the party tonight, and neither does the choice of your lunch affect the choice of your clothes.

- F2 - Second, people don't seem to work out detailed scenarios of the future about all possible consequences of their choice. When choosing your lunch, you are probably thinking about how much hunger you have and what lunch would best satisfy that hunger. You are, probably, not thinking about what this particular lunch would do for your health in two weeks. That would simply be too much.

- F3 - Third, when making a decision, one tends to look at relevant aspects only. Buying lunch will probably focus your attention to various aspects of available lunches, and divert your attention from other domains, such as music or clothes. Instead of taking into account all possible information, you concentrate on the specific domain of choice.

These aspects of decision-making are incorporated in theories of bounded rationality. In most situations we can discriminate only few factors that influence our choice. These are the factors that seem important to us. When you buy your lunch, you will probably look at the size, price and tastefulness of different lunches before making a decision. Other factors, such as the percentage of vitamin B_6 or the exact weight of the lunches will be left unconsidered. If we wouldn't do such a thing, it would be practically impossible to make as many decisions as we do in everyday life. The human species wouldn't have come this far using such an elaborate and complex mechanism. The complexity of the world makes optimization a very costly and difficult process, and genuine optima are most of the time simply not computable within feasible limits of human cognitive effort.

2.2.3. Fast and frugal heuristics

Then how do humans make decisions? Humans can make good decisions in an effective and efficient way, as proven by our everyday life. So there must be a mechanism that provides us with the tools to make these decisions. Such a mechanism must be realistic – thus able to cope with limited time and limited resources – and it must be reliable. Otherwise it would not be able to explain why humans can make effective and efficient decisions.

Todd and Gigerenzer propose a solution: fast and frugal heuristics (Todd & Gigerenzer, 1999; 2000). Fast and frugal heuristics are simple rules that we use for making decisions with realistic mental resources. Fast and frugal heuristics can be as accurate as strategies that use all available information and expensive computation (Todd & Gigerenzer, 1999; 2000). This mechanism can deal with multiple alternatives (F1), and can especially be used to make choices between simultaneously available alternatives. The search for information about the different options is limited (F2 and F3), rather than the search for the options themselves. These heuristics are fast, because they do not involve much computation. They are frugal because they search only for some of the available information. Examples of fast and frugal heuristics are yes/no decision trees, or one-reason decision-making (choosing an alternative based on one aspect).

Being fast and frugal are important aspects of simple heuristics, because it makes them realistic (for practical reasons I will use the terms simple heuristics and fast and frugal heuristics as being identical). Such heuristics can cope with real-life situations with limited time and limited knowledge. But are they reliable? Can such heuristics explain how it is possible that humans make fast and smart decisions? Todd and Gigerenzer show several times that simple heuristics can perform almost as well as complicated and time-consuming algorithms or mathematical techniques (Todd & Gigerenzer, 1999). In some occasions the computational techniques are even outperformed by simple heuristics. This shows that such a mechanism is not only realistic, but also reliable.

To show the workings of fast and frugal heuristics, I will describe a brief example adopted from (Todd & Gigerenzer, 1999). Wild Norwegian rats have an eating habit called *neophobia*, i.e. a reluctance to eat foods that they do not recognize (Bartlett, 1932). Recognition can be based on the rat's own experience, but also from smelling foods on the breath of other rats (Galef, 1987; Galef et al., 1990). This is a smart heuristic, because every food that a rat has eaten during his life, hasn't killed it (Revusky & Bedarf, 1967). This heuristic for food recognition is even followed when the rat has smelled it from the breath of a sick rat. It is important to see the power of such a simple heuristic: following this rule, rats can survive (except when they disobey the illness information). It is a rule that does not involve complex and time-consuming calculation, but instead uses an evolutionary shaped system in the rat that works fast: recognition.

Fast and frugal heuristics have incorporated the three mentioned aspects of decision making. They are also realistic and reliable. In this work, I will investigate decision-making in communication. Generally, the alternatives for interpretation or production of utterances are encountered simultaneously in communication. Fast and frugal heuristics can deal with such alternatives in decision-making processes, and they can provide us with a realistic and reliable mechanism for decision making. So, I will have to find a suitable form for fast and frugal heuristics in the domain of language. To use the principle of fast and frugal heuristics, I need to have a theoretical framework for communication in which these heuristics can be incorporated. This theoretical framework can be found in optimality theory. In the following Section 2.3 I will describe this theory and show how we can use it to come to a solution.

2.3. Optimality Theory

2.3.1. Introduction

In this section I will first explain the basic ideas and workings of Optimality Theory (OT). The previous Section 2.2 provided us with an interesting mechanism that can explain the ease with which humans can make complex decisions in a limited environment. I will incorporate the idea of fast and frugal heuristics within OT, by showing several parallels between ideas in OT and fast and frugal heuristics.

Optimality Theory describes the grammar of a language as a set of conflicting constraints that have to be resolved for each formulation (Prince & Smolensky, 1993; Archangeli & Langendoen, 1997) and interpretation (Hendriks & de Hoop, 2001) of utterances in discourse. In certain situations the constraints will conflict and this conflict has to be resolved. This is done based on an ordering of the different constraints at stake.

Prince and Smolensky (Prince & Smolensky, 1993) have first proposed Optimality Theory in the field of phonology. Their basic idea is that grammar consists of a set of universal constraints on well-formedness. These constraints are the building blocks of a grammar. Different grammars have a different ordering of constraints. An important aspect is that some constraints are highly conflicting. Because of this conflicting nature, the constraints will be violated often in the actual, everyday forms of language. The constraints in phonology are ordered in a strict dominance hierarchy in which every constraint has absolute priority over all lower-ranked constraints. A grammar is in fact the ranking of universal constraints into a strict constraint hierarchy. In phonology and syntax these aspects of OT have been accepted widely. Recently these aspects have also been suggested in semantics and pragmatics (Blutner, 2000; Dekker & van Rooy, 2000; Zeevat, 2000; Hendriks & de Hoop, 2001).

In order to explain OT in the domain of interpretation, I will start with a description of the blocking and triggering effects. Grice (Grice, 1975) has first explained these effects with his maxims of conversation. These maxims have been reformulated into the Q- and I-principle (Horn, 1984), which I will also discuss. These principles can be seen as pragmatic constraints in OT, using a bi-directional viewpoint as Blutner proposes (Blutner, 2000). This will then allow us to see the parallels with the mechanism of simple heuristics from the previous Section 2.2.

2.3.2. Blocking and triggering

In everyday language constraints interact. Two patterns of constraint interaction seem to appear often. These patterns are known as blocking and triggering. Blocking occurs when a specific condition limits the scope of an otherwise broadly applicable generalization. In the field of pragmatics this is also known as a marked situation. A marked situation is, simply said, an unusual situation. On the other hand there is the unmarked situation, the stereotypical situation.

The general tendency seems to be that 'unmarked forms tend to be used for unmarked situations and marked forms for marked situations' (Horn 1984: 26). This tendency is known as *the division of pragmatic labor* (Horn, 1984). Let's clarify this with an example (from Blutner, 2000):

(1) a. Black Bart killed the sheriff.
b. Black Bart caused the sheriff to die.

Sentence (1a) is clearly the unmarked expression. This is the stereotypical way of formulating that Black Bart killed the sheriff in a typical Wild West gunfight. This brings us to sentence (1b): when you read this sentence, you automatically assume that there was something unusual about the way Black Bart killed the sheriff, e.g. by causing his gun to backfire by stuffing his gun with cotton. So the unusualness of the expression – taking into account your knowledge about the English language, Wild West bandits and sheriffs – forces you to interpret the sentence in a different way, i.e. to assume that there is something unusual about the way Black Bart killed the sheriff. This is called blocking, because the standard, unmarked meaning (to kill in a gunfight) is blocked by the special, marked meaning (to kill by causing his gun to backfire).

Triggering, on the other hand, is the opposite of blocking. So, sentence (1b) is triggered by the unusualness of the situation. In a normal situation, this sentence would have been blocked. The triggering and blocking principles have effect in both formulation and interpretation of utterances. By uttering sentence (1b) you will try to communicate that something special is happening. On the other hand, by hearing (1b) you will assume that something special is happening. So (1b) can also be seen as a sentence that triggers the special meaning (to kill by causing his gun to backfire), which blocks the usual meaning (to kill in a gunfight). The process of triggering and blocking can thus be seen as being symmetrical. When a sentence triggers a certain meaning, it thus in effect blocks another meaning and the other way around. As we will see below, blocking and triggering effects can be explained within bi-directional OT.

2.3.3. Q- and I-principle

Grice (Grice, 1975), has explained these blocking and triggering effects by his maxims of conversation. Here is a brief description of his maxims of conversation (Grice, 1975):

Quality: try to make your contribution one that is true
- Do not say what you believe to be false
- Do not say that for which you lack evidence
Quantity:
- Make your contribution as informative as required
- Do not make your contribution more informative than is required
Relation: be relevant
Manner: be perspicuous

- Avoid obscurity of expression
- Avoid ambiguity
- Be brief
- Be orderly

He introduced pragmatics into the field of linguistics with these maxims. Language use can be seen as a special kind of cooperative behavior (Grice, 1975). Grice's principles have been reduced – or better said: reformulated – in the Q- and I-principles (Atlas & Levinson, 1981; Horn, 1984). These principles govern our daily communication processes as illustrated in the Black Bart example. They are stated below and clarified in Table 2.1

Q-principle: say as much as possible to fulfil your communication goals.
The speaker has to be as informative as possible. The speaker's efforts are maximized by this principle.
The hearer, on the other hand, will have little trouble understanding the utterances of the speaker, since the speaker is so informative. The Q-principle thus minimizes the efforts of the hearer.

I-principle: say no more than necessary to fulfil your communication goals.
The hearer has to extract as much information from the speaker's utterance as possible. His efforts are maximized.
The speaker minimized his efforts, because the hearer will do his very best to try to understand what the speaker is saying.

	Speaker	**Hearer**
Q-principle Say as much as possible	Maximize efforts (to facilitate the hearer's understanding)	Minimize efforts (since speaker is so informative)
I-principle Say no more than necessary	Minimize efforts (such as time, articulation, etc)	Maximize efforts (extract as much information as possible from the utterance)

Table 2.1 The Q- and I-principles

There is another principle that is worth our consideration here: the principle of linguistic economy. The principle of linguistic economy says that the speaker is maximizing profits by restricting resources, such as time, articulatory effort, memory and attention. The hearer is seeking to maximize his understanding extracting as much information as possible from what is said, while minimizing his cognitive effort and economizing processing cost (ter Meulen, 2000).

Principle of linguistic economy: speaker and listener try to maximize understanding, while minimizing their efforts.

This principle can be seen as a combination of the Q- and I-principles, although the Q- and I-principles are a more detailed description and incorporate the cooperation between the hearer and the speaker explicitly. I will return to this subject shortly in Section 2.4. Because the Q- and I-principles are more detailed, I will concentrate on these principles here and regard the principle of linguistic economy as further evidence for optimalization processes in natural language processing for both speaker and hearer. Blutner has formulated a bi-directional form of OT using the Q- and I-principles. In order to persue this idea in more depth, I will now first explain Optimality Theory and I will then return to the idea of bi-directional OT.

2.3.4. Structure of Optimality Theory

In OT there are three formal components: the GENerator, the EVALuator and the set of ranked CONstraints. The constraints have different strengths, meaning that one constraint can dominate another. These components work as follows: consider a certain input A. For this input, GEN creates a candidate set of possible outputs, B_i. From this candidate set, EVAL selects the optimal output B using the different ranked constraints from CON. This then, is the output that resolves the conflicts between the different active constraints in an optimal way. The optimal way is determined not by satisfying the most constraints, but by satisfying the strongest possible constraint.

The three formal components have been discussed in phonology (see e.g. Prince & Smolensky, 1993). To clarify the working of the three components, I will discuss a brief example from Prince and Smolensky (Prince & Smolensky, 1997). It's an example from OT syntax. Take the English sentence *it rains*. This is a normal and well-formed English sentence. This means it must be the optimal outcome of the process between GEN, EVAL and CON. So *it rains* is the optimal output B. It was selected from a candidate set of possible outputs, B_I, generated by GEN. A simple possible set of sentences could be {*rains, it rains*}. Note here that the possible set of sentences is generally considered to be an infinite set. This is because the desired input A (the message) can be formulated in endlessly different ways in the language. Other constraints are generally assumed to diminish the set of possible sentences. Therefore we will consider only this small set here. In this case we can distinguish at least two constraints from CON:

CONTRIBUTE: all words have to contribute to meaning
SUBJECT: all sentences need to have a subject

As we can see the element *it* has no clear meaning in this sentence. But this element solves the constraint SUBJECT. So, the sentence *it rains* has been selected from the set of possible sentences, because it solves the constraint SUBJECT and SUBJECT is apparently stronger than CONTRIBUTE in the English language.

Remember that in phonology and syntax there is a dominance hierarchy of constraints. This is generally also assumed for semantics. Constraints are 'soft', in a way that they can be violated. An output that has a violation of a higher-ranked constraint can never win over outputs that have several violations of lower-ranked constraints (Prince & Smolensky, 1993; Blutner, 2000; Hendriks & de Hoop, 2001). The ranking of constraints is language particular.

Prince and Smolensky (Prince & Smolensky, 1993) formulated the Panini theorem, which intuitively says the following: if a more specific constraint is lower-ranked than a general constraint, then it will be over-ruled by the higher-ranked constraint with which it conflicts. So, if a specific constraint is to have any effect, it needs to be higher-ranked. The Panini theorem thus allows us to spot the ranking of certain constraints in given situations. When we take a look at our example sentence *it rains*, the Panini theorem is what makes it possible to see for us that SUBJECT has a higher ranking than CONTRIBUTE: SUBJECT > CONTRIBUTE.

Generally in OT this is shown in a tableau. In a tableau the constraints are ranked across the top, going from the highest ranked on the left to the lowest ranked on the right. An asterisk (*) shows a violation of a constraint, where an exclamation point (!) shows a fatal violation, i.e. a violation that eliminates a candidate completely. The little hand (☞) marks the optimal candidate (Archangeli & Langendoen, 1997). For our example sentence *it rains* the tableau is shown in Table 2.2.

	SUBJECT	CONTRIBUTE
☞ *It rains*		*
Rains	*!	

Table 2.2 Tableau for it rains

2.3.5. Bi-directional OT

The need for bi-directional OT will become clearer if we concentrate on the two different roles a communicator has in a conversation. The first role is that of the speaker, producing utterances. The second role is that of the hearer, interpreting utterances. Now, the role of the speaker has been considered purely from a syntactic point of view: the speaker wants to communicate a certain semantic input to the hearer. OT syntax optimizes the syntactic structure (the surface structure) with respect to this semantic input (the underlying structure). The role of the hearer is then seen from a semantic point of view. OT semantics optimizes the semantic structure with respect to the syntactic structure (Hendriks & de Hoop, 2001).

What this basically means, is that the speaker wishes to communicate a certain meaning and that there are different possible utterances that will reflect this meaning. The optimal utterance for the desired meaning is selected by OT syntax. The hearer on the other hand, interprets the utterance with respect to the different possible meanings for this utterance. Here the optimal meaning is selected by OT semantics. What Blutner actually says, is that this is strange: why would one person constantly switch between these roles? Both the hearer and speaker roles are available for this person! A speaker can also use her hearer role to narrow down the best utterance for a certain meaning, since she can interpret what she will say herself too. This also goes the other way around: a hearer can interpret the utterance not only with respect to the possible meanings of this utterance, but also with respect to the possible utterances the speaker had available.

Blutner therefore slightly reformulates the Q- and I-principle, saying that the I-principle seeks to select the most coherent interpretation, and the Q-principle acts as a blocking mechanism and blocks all the outputs that can be derived more economically from an alternative input (Blutner 2000). This way the Gricean framework can be grasped by a bi-directional OT framework, taking into account both the hearer and the speaker perspective. This bi-directional viewpoint is also supported by other literature such as (Dekker & van Rooy, 2000; Zeevat, 2000).

Remember the Black Bart example. We should now be able to explain the blocking and triggering effects in this example. According to Blutner (Blutner, 2000) there are two conflicting constraints at work here:

C: (semantic/pragmatic) prefer coherent and informative expressions
F: (syntactic) use standard, usual forms when possible

F is a constraint on form, while C is a constraint on meaning. A standard and usual form is a form that is frequently used. *Cause to die* is a non-standard and unusual form that is not frequently used. I will display a tableau here, adapted from (Blutner, 2000). Since we use a bi-directional form of OT, we will have to show both the forms (syntactic structures) as well as the interpretations (semantic meaning) of the two options. Blutner adds the little arc (➳) to indicate the optimal semantic candidate (the optimal interpretation for the hearer) and uses the little hand (☞) to indicate the optimal syntactic candidate (the best form to be used for the speaker). Table 2.3 shows the tableau for this example.

Forms		F	C		F	C
killed	➳ ☞					*
caused to die		*		➳ ☞	*	*
		shot			*stuffed gun with cotton*	
		Interpretations				

Table 2.3 Tableau for the Black Bart example

This is a quite complex tableau, so I will try to clarify it. We see that the expression *caused to die* violates F with respect to both interpretations and that the meaning *stuffed gun with cotton* violates C with respect to both forms.

If a hearer would hear the expression *killed* both interpretations are available, but the interpretation *shot* is optimal, because it is more economical. If the speaker wished to communicate the meaning *shot*, both forms are available, but the form *killed* will be selected, since it is more usual, more frequent.

If the hearer hears the expression *caused to die* the only available interpretation is *stuffed gun with cotton,* since the unmarked form *killed* blocks the meaning *shot*. This input-output pair has already been coupled as being the optimal input-output pair. It works analogously for the speaker. If the speaker wishes to communicate the meaning *stuffed gun with cotton* the only available form is *caused to die*, since here *killed* is being blocked.

2.3.6. Game Theory

The ranking and judging of syntactic and semantic structures in bi-directional OT has a structure that resembles certain aspects of game theory. I will now discuss these parallels shortly using the notion of game theory. Dekker and van Rooy (Dekker & van Rooy, 2000) point out several parallelisms between some notions studied in OT and game theory. Optimality theoretic interpretation can be seen as an interpretation game. The solution concept for this interpretation game is optimality. Optimality, in turn, can be defined as a Nash Equilibrium of the interpretation game. A Nash Equilibrium is a profile in which each player's action is a best response to the choices of the other players in that profile. This is like the Q- and I-principles as described above: both players (speaker and hearer) try to maximize mutual understanding and to minimize efforts, with respect to eachother. (For a more detailed description see Dekker & van Rooy, 2000).

In such a interpretation game the players have different roles. One can be a speaker, while the other is a hearer. The speaker wants to communicate a certain meaning, in pursuing a certain goal. As action-set he has the set of possible representations. The chosen representation will be uttered (or typed) and will be received by the hearer. This person has to assign a suitable meaning to the representation chosen by the speaker. As action-set he has the set of possible meanings. Note that in an experimental setting these action-sets can be restricted to finite controllable sets. This will be discussed in more detail in Chapter 4. Natural language communication can thus be seen as an interpretation game, where players have interchanging roles of speaker and hearer. This is basically the same viewpoint that Blutners bi-directional OT takes: here communicators interchange their roles of speaker and hearer and make use of their interpretation and production skills in both roles.

2.3.7. Simple Heuristics

We have at this point discussed OT and the bi-directional form that Blutner has provided. We have also seen a solution concept for OT: optimality as a Nash Equilibrium. I wish to return to the subject of fast and frugal heuristics now, in order to show the parallels between the theoretical frameworks and integrate them.

We have seen an OT analysis of the sentence *it rains*. In this sentence we distinguished two constraints SUBJECT and CONTRIBUTE and we inferred that SUBJECT has a higher ranking than CONTRIBUTE in English. To show the parallels between bi-directional OT and fast and frugal heuristics, I will compare this OT example to the example of the Norway rats. If we look at the latter example closely, we can distinguish two heuristic rules[1] that influence the rat's food choice: the first heuristic rule is that rats will prefer food that they recognize. The second heuristic rule is that rats will not prefer food of which they have illness information, meaning that they have smelled the food on the breath of a sick rat. For reasons that will become obvious I will write these heuristic rules down as follows:

RECOGNITION:	prefer food that has been recognized
ILLNESS:	do not prefer food of which you have illness information.

Normally, rats will prefer food that they recognize from having tasted them or from having smelled them on the breath of another rat (Galef, 1987; Galef et al., 1990), following the RECOGNITION rule. But if the rat is confronted with food of which it has illness information (do not eat) and recognition (do eat), RECOGNITION will dominate ILLNESS, that is: RECOGNITION > ILLNESS. I hope that at this point the strong resemblance between heuristic rules and constraints is obvious. They both have similar structures, building complex behavior on simple ranked rules (heuristic rules or constraints) and using this ranking in situations where rules are in conflict.

Further evidence for such cognitive systems based on ranked and sometimes conflicting rules can also be found in the visual field. Take for example Petter's Illusion in Figure 2.1 (adopted from Todd & Gigerenzer, 1999). In this figure, some people see the foil going through the referee. (Also some people see the legs of the left fencer in front of the fence.) There are several visual cues that are being used by the human visual system for depth perception, e.g. object OVERLAP (closer objects sometimes partially obscure farther objects); HEIGHT IN THE PLANE (objects that are farther away are closer to the horizon); object SIZE (objects that are farther away appear smaller); object COHERENCE (a segment of a figure that has the same appearance will be regarded as one object), etc. Here we see another parallel: cues in the visual systems can be compared to constraints in OT, or heuristic rules in decision-making. The illusion is caused by the domination of one cue over others: here OVERLAP can't be used to interpret the difference between the foil and the referee because they have the same color, so COHERENCE causes some people to see the foil and

[1] In Simple Heuristics That Make Us Smart, the authors insufficiently discriminate between heuristic rules and heuristic strategies. They use the term heuristic both for rules and for strategies. I propose to make a clear distinction between the two. Heuristic rules are what I compare to constraints: simple rules that say what to do. Heuristic strategies are strategies that are based upon a system of ranked heuristic rules. In my viewpoint OT can be regarded as a heuristic strategy, using constraints as heuristic rules and optimality as a solution concept.

the referee as one. This shows that OVERLAP > COHERENCE (if the overlap is discriminable) and COHERENCE > SIZE, HEIGHT IN THE PLANE (for the people who see the foil going through the referee).

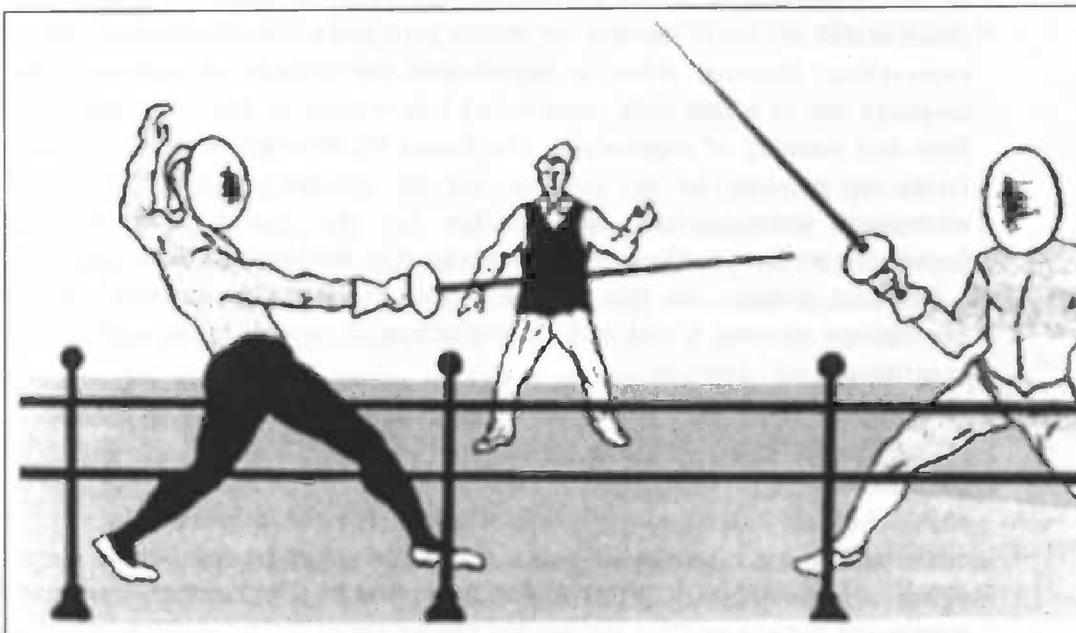

Figure 2.1 Petter's illusion

So, the basic idea is that in natural language interpretation there is a system of ranked and sometimes conflicting constraints. This system is collected in the theoretical framework of OT. I have shown parallels in other cognitive systems that support such a system in the field of natural language interpretation. To resolve the conflict between these constraints, we need a strategy. The strategy we have set out here is bi-directional OT, with optimality defined as a Nash Equilibrium as a solution concept.

2.4. Referring Expressions

2.4.1. Introduction

In this project I wish to investigate the role of bounded rationality in communication. I will do this by investigating the use of nominal referring expressions. Nominal referring expressions are a clear example of natural language use in which both speaker and hearer have to determine the optimal form and meaning of expressions. The hearer needs to know which alternative forms are possible for the speaker and the speaker needs to know which alternative interpretations are possible for the hearer. The theoretical framework we have explored so far explains this: Blutner's bi-directional OT as a heuristic strategy. In this section I will explain what nominal referring expressions are, and I will make a distinction in several types useful for the experiment I will perform.

Referring expressions are linguistic forms that are used to pick out an entity in the context (Dale, 1992). The term entity here has to be regarded not only as a physical object in the real world, e.g. an apple, but also as possible subjects for conversation, e.g. a movie, love, etc. So an entity can be any object or entity, either real or imaginary, either physical or abstract. The context must also be considered not too strictly: the context can be either direct and physically near, or it can be indirect and not present. The context of a conversation, including the conversation itself up till the moment of consideration, is generally referred to as the discourse (Roberts, 1999). The discourse includes the physical and imaginary environment of the conversation as well as the entities mentioned earlier in the conversation. I will return to the subject of discourse in Section 2.5.

2.4.2. Anaphora

Anaphora are a special form of referring expressions: anaphoric referring expressions. The entity picked out by an anaphoric expression can be determined only by making use of contextual information, and not from the content of the form itself (Dale, 1992; Reinhart, 1999a). An anaphor thus lacks clear independent reference and picks up its reference through connection with other linguistic elements. Normally this is the case when two nominal expressions are assigned the same referential value or range: an element of the discourse. Generally anaphora tend to be abbreviated linguistic forms, meaning that they are shorter and thus more economical. Let's look at a short example to clarify the use of anaphoric referring expressions:

(2a) *I had a cheese sandwich today for lunch.*
(b) *I liked it.*

In this context, the use of the pronominal anaphor *it* in (2b) is readily interpreted as referring to the cheese sandwich I had for lunch today. Even if I would not have uttered sentence (2a), but instead was sitting at a lunch-table with my colleagues and I had just finished eating my cheese sandwich, *it* would easily be interpreted the same. As mentioned above, an important characteristic of anaphora is that they do not refer directly to an entity, but require other elements of the discourse for their interpretation. The elements that can be used for interpretation are all part of the common ground, as we will see later in Section 2.5. The interpretation of anaphoric expression using the context is called anaphora resolution (Reinhart, 1999b).

2.4.3. Anaphora Resolution

This process of finding an interpretation of an anaphoric expression is widespread in our everyday life. The entity that the anaphoric expression is referring to is called the referent. When people are talking, they make frequent use of referring expressions and anaphora. Using these expressions is more economical (linguistic economy), but it also makes clear that the subject is already part of the discourse. In this way coherence is generated in the discourse. Pronouns are a good example of this phenomenon: referring to a person or animal by saying *it* or *she* is much shorter than using a descriptive expression, such as *The girl with the red hair that lives next-door*. This last form might very well be used to introduce the subject into the conversation. These expressions are called nominal referring expressions, since they refer to an entity in the discourse and appear in the form of a noun phrase. Note again that an entity does not have to be real or physical.

Anaphoric resolution is not a slow conscious process, but a fast unconscious mechanism. We use it numerous times during a conversation. As I discussed in Section 2.2, it is not likely that such a process is complex and based on vast computation. The restricted capacities of the human cognitive system would simply not allow for such a complex process. It is more likely that the use of nominal referring expressions can be explained by the theoretical framework that we have established so far. Interpreting nominal referring is a process in which principles such as the Q- and I-principle have important roles. Both speakers and listeners act according to Blutner's bi-directional OT, in order to produce and interpret nominal referring expressions in an optimal way.

2.4.4. Types of referring expressions

Givon has distinguished several different expressions with respect to their discourse function (Givon, 1983). He presents a scale in the coding of topic accessibility. It refers to the availability of expressions. If two persons are talking about their colleague, and one would say: *she has such a nice suit*, the word *she* is easily interpreted as referring to their mutual colleague, since they were just discussing her: their mutual colleague is an easily accessible item in the discourse. On the other hand, if one of them wants to refer to their lunch all of a sudden, this will be a non-salient item and it will need to be introduced as a topic. Givon has formulated the following principle to account for this effect:

> The more disruptive, surprising, discontinuous or hard to process a topic is, the more *coding material* must be assigned to it. (Givon, 1983: 18, his italics).

Basically, this principle accounts for the phenomonon that participants of a conversation can easily speak about the current topic, but will have to spend more energy if they wish to change the topic. From Givon's scale we can conclude the following (see Givon for the full scale):

More salient topics
↑ Null anaphora
Unstressed pronouns
Stressed pronouns
Definite noun phrases
↓ Indefinite noun phrases
Less salient topics

The important thing to see here is that null anaphora can be used for the most accessible topics, followed by definite noun phrases, while indefinite noun phrases will be used for the least accessible topics. For the purpose of my experiment I will distinguish three types of referring expressions based on Givon's scale.

INDEFINITE: Indefinite noun phrases: *a blue dot is left of the green circle.*
These are noun phrases beginning with the indefinite determiner *a*. Generally an indefinite noun phrase would be used to refer to a non-accessible object.

DEFINITE: Definite noun phrases: *the blue dot is left of the green circle.*
These are noun phrases beginning with the definite determiner *the*. Normally a definite noun phrase would be used to talk about an entity that is more accessible in the discourse.

NULL: Null anaphora: *the blue one is left of the green circle.*
These are anaphoric expressions that are in a sense 'empty'. Please note that in Dutch null anaphora have a different structure. In Dutch the adjective *blauwe* (blue) functions as a noun: *de blauwe* means *the blue one*. This means that in Dutch the use of a null anaphor has a clear economical advantage over a definite expression. The referent of the expression needs to be highly accessible, in order to make the reference clear.

Within each of these types a second distinction can be made with respect to the amount of descriptive words used in the phrase. In our experiment there is a maximum of three adjectives to describe an object. These are global position, size and color. I will give examples here for DEFINITE expressions:

3 ADJ:	*the leftmost large blue circle*
2 ADJ:	*the large blue circle*
	the topmost blue circle
	the topmost large circle
1 ADJ:	*the blue circle*
0 ADJ:	*the circle*

2.5. Common Ground

2.5.1. Introduction

In this section I will concentrate on the discussion of context. It is clear that speakers and hearers form and interpret each utterance against a background of information, to which we usually refer as context or discourse. These terms however, seem to remain rather vague. In Section 2.4 above, I simply described the discourse as the context of a conversation, including the conversation up till the moment of consideration (Roberts, 1999).

The pragmatic OT constraints from bi-directional OT cannot be understood without a concept of context. Participants of a conversation can only cooperate in communication if they understand the messages that are being exchanged and can make their own messages understandable. Even when a conversation starts, participants need to have a basis of shared information, otherwise they would never understand each other. The language they speak and the knowledge they have of the world of course largely provide this basis. I will refer to this knowledge as background knowledge.

2.5.2. Background knowledge

Background knowledge is the information that you have acquired because you are member of a specific community. It is the information you have learned because of your similar background or education (Lee, 2001). It is the knowledge that belongs to your culture and society, your form of life so to say. Knowledge of the language you speak is a good example of background knowledge: if the person you are communicating with speaks the same language, you can conclude that he or she speaks and interprets according to the same rules as you do.

There is a note to be placed here: there are different forms of life, on different levels. For example: one form of life is the country you live in. Another might be your home, your family and their specific ways. Another might be the field you work in, e.g. the field of linguistics. All these forms of life have their own specific 'rules', behaviors, words, theories and facts.

Basically we can say that each conversation starts with a collection of shared information, depending on the form of life. When the conversation takes place, more information is shared. The participants in fact accumulate this information by adding to it with each utterance or speech act (Clark, 1992). This information in turn can also be used as a common background for reference. Hence the term common ground.

Up till now I used the term *information* to describe what participants of a conversation have (deliberately) shared in their common ground. Clark describes common ground as the sum of the mutual (or common) knowledge, mutual beliefs and mutual presuppositions of the participants (Clark, 1992). Note here that the term *mutual* can be substituted with the term *common*, meaning that all participants must know that all participants know, etc. See also Paragraph 2.5.4.

So, the former description of the common ground is the *information* that has been accumulated during discourse. Remember that the common ground includes the physical and imaginary environment of the conversation as well as the entities mentioned earlier in during discourse. Everything that has been mentioned or agreed upon in a certain conversation, is part of the common ground. Next I will discuss the difference between knowledge and belief, then I will define the group notions of common knowledge and belief.

2.5.3. Information as knowledge or belief

The difference between knowledge and belief is one of much discussion. Different scholars tend to use the terms differently and the line between them can be thin. Let's take a look at some examples: you may *believe* that you can have a cheese sandwich for lunch today at the deli, because you had that yesterday and the deli will probably serve the same today. Or you may *know* that you had a cheese sandwich for lunch, because you already ate it. I do not wish to solve the discussion here, or to present a viable definition of either knowledge or belief. Rather, I will adopt Clark's use of the term common ground, as a general term that covers mutual knowledge, mutual beliefs and other mutual attitudes.

Clark also uses the term *know* as a more general term, which might well be replaced by other mental attitudes, such as believing. I will now illustrate the difference between knowing and believing, using a short example. Then I will discuss some formal aspects of common knowledge or belief.

Remember the example about Black Bart and the sheriff. Let's say that

q = Black Bart stuffed the sheriff's gun with cotton

Let's also say that Black Bart in fact did stuff the sheriff's gun with cotton and let's say that Black Bart was aware of his gun-stuffing actions. In that case it holds to say:

$K_B q$ = Black Bart knows q

Where K_X stands for: agent X knows. Now let's say that no one has seen Black Bart stuff the sheriff's gun, and Black Bart has told no one about his sabotage action. So, Black Bart is the only person in the world that *knows* q. Other persons may have suspicions about what Black Bart did, but they will not know if q is true. Here we see the difference between *knowing* and *believing*. While Black Bart can know q, other persons can only believe q, because they cannot check whether q is true.

Now let's also say that the sheriff has heard some stories about Black Bart's immoral nature and that he has reason to believe that Black Bart has stuffed his gun with cotton:

$B_S q$ = the sheriff believes q

What is important for my experiment, is not so much the difference between knowing and believing. Following Clark I can use a general term, which is replaceable with other terms concerning mental attitudes. Since in my experiment participants will receive much second-hand information, I will use the term *believe* as the general term. This term is more justifiable.

2.5.4. Common knowledge and belief

For information to become part of the common ground, it is necessary to be mutual or common belief. A participant must be convinced that their communication partner also has the information at stake. Furthermore the participant must also be convinced that their partner is convinced that they have the information at stake, and so on. I will continue our example, to show what it takes to become common belief. There were two agents so far (I will replace *know* for *belief* from now on), Black Bart and the sheriff. Our logic statements about them are:

$B_B q$ = Black Bart believes q
$B_S q$ = the sheriff believes q

For simplicity we will assume that there is no one else in town, so our two agents are the only agents at consideration here. If all agents believe q, everyone will believe q:

$B_B q \wedge B_S q \rightarrow Eq$ = everyone believes that q

But does this mean that all agents believe that all agents believe that q? Would it mean that both Black Bart and the sheriff believe that they themselves and the other believe that q? This is not the case: how should Black Bart believe that the sheriff believes q? He might believe that the sheriff does not trust him, but how should he suspect the sheriff to see through his plans? To become common belief, the knowledge needs to be shared. In this example: Black Bart should tell the sheriff that he stuffed his gun with cotton. The sheriff would then believe q. But does Black Bart believe at this point that the sheriff believes that Black Bart believes that q? In fact, if Black Bart believes that the sheriff does not trust him, Black Bart does not believe that the sheriff believes that Black Bart believes q. It is probably clear that we could continue this example forever.

So, for q to be common belief, everyone has to believe q, and everyone has to believe that everyone believes q, etc ad infinitum. If we could write this in a logical formula, it would look something like this:

$$Cq = Eq \wedge EEq \wedge EEEq \wedge \ldots \text{ ad infinitum}$$

(Unfortunately, this is not allowed in epistemic logic.) In our example this means the following: it is common belief that Black Bart stuffed the sheriffs gun with cotton, if both Black Bart and the sheriff (everyone) believe that Black Bart stuffed the sheriffs gun with cotton, and everyone believes that everyone believes that that Black Bart stuffed the sheriffs gun with cotton, etc. Usually such common beliefs are present from the beginning of a conversation or arise during and from communication. A good description of the notion of common knowledge in epistemic logic can be found in Van der Hoek & Verbrugge (2002).

2.5.5. Anchoring

Now we have a basic description of a conversation: participants start with background knowledge, which is connected to the appropriate form of life. At the start of a conversation this is the basis you will work with. From here on participants will add information to the common ground and use this for reference. The specific environment, e.g. completing a task, in which the conversation is taking place, can add specific information to the common ground. When you are going out for lunch with a colleague, saying *they make great sandwiches here* will be a non-ambiguous utterance.

What speakers must do in order to refer properly is the following: the speaker must have good reason to believe that his utterance will introduce the subject to the common ground. To become part of the common ground, a referent must at least be anchored to something that is already part of the common ground (Clark, 1992). This means that participants need not have a common ground in which the referent already exists. Participants need to be able to form or expand their common ground based on their background knowledge and on the speech act that took place. In this anchoring principle we find a condition for expanding the common ground, which has to be satisfied in my experiment.

2.6. ACT-R

To set up the experiment I have chosen ACT-R as a modeling environment. Matessa (Matessa, 2000a; 2000b) developed a graph completion task. In this task, participants have to communicate with each other, in order to solve a puzzle. I have chosen to use this model and experimental software, since it had already been developed. In this section I will discuss the cognitive theory of ACT-R, since I use ACT-R as a modeling environment. In Chapter 4 I will discuss the properties of this experiment and several other reasons that made this experiment suitable.

ACT-R is a very suitable modeling environment, in which it is possible to create a user model. Matessa has created such a user model as a communication partner in his experiment. This allows us to use a model of a participant, instead of a real human participant. This model can interact with another participant, as if there were two human participants. Basically a human participant will play against a communication model, allowing us to assign the linguistic variation to the human participant. Such a model cannot be regarded as a serious attempt to create a working model of a communicating human, such as in the well-known Turing test.

ACT-R is a computational theory of human cognition. It assumes there are two types of knowledge: declarative knowledge and procedural knowledge (Anderson & Lebiere, 1998). Declarative knowledge can be described as the knowledge you are aware of and can usually explain to others. For example: simple addition facts like *three plus four is seven*. Procedural knowledge is knowledge that we display in our behavior, but that we are unaware of. It is the knowledge that allows you to remember how to drive a car, or how to solve an addition problem. Procedural knowledge basically specifies how to use declarative knowledge to solve certain problems (Anderson & Lebiere, 1998).

In ACT-R these two forms of knowledge have been combined in a production system. Procedural rules (representing procedural knowledge) act on declarative chunks (representing facts in memory). The production rules and the chunks are considered as the symbolic level, representing the types of knowledge. Every production rule and chunk has several attributes that can vary during a run of the production system. These attributes influence the probability of retrieval, the time it takes to retrieve a fact and the strength of production rules. The distinction between declarative and procedural, and between symbolic and subsymbolic is represented Table 2.4.

	Declarative	**Procedural**
Symbolic	Facts	Rules
Subsymbolic	Activation	Reliability

Table 2.4 ACT-R

Every rule has a set of conditions that need to be true in order for the rule to fire. The production rules work in a serial fashion: each of the rules is considered and the rule that matches will fire. When two or more rules have satisfied conditions, the rule that has the highest expected gain will fire. Generally a rule will retrieve a chunk from memory and perform a certain operation on this chunk. Chunks need to have a certain activation to be retrieved from memory. ACT-R has variable threshold values for chunk-retrieval and a chunk has to have an activation above this threshold to be retrieved from memory.

New chunks can be added to the declarative memory when a problem is solved. The theory stipulates that there are only two sources of new chunks: from perception and from completed goals. For example: for an ACT-R model about addition problems, there are two ways of solving an addition problem: by computing the answer (using production rules to retrieve chunks and combining these to form the answer), of by retrieving the answer from memory. Initially the model will compute answers in a procedural way. When the answer to a certain problem is computed, it will be stored in the goal. From now on, the answer will be available as a fact in the declarative memory. Whether or not the new chunk will be used is a matter of activation. In this way an ACT-R production system is capable of learning new facts and achieving a faster response. (For an in depth discussion on ACT-R see (Anderson & Lebiere, 1998).)

2.7. Chapter Summary

In this chapter I have laid down the theoretical framework I need for my research and experiments. I discussed bounded rationality, a theory that incorporates the limitations of the human cognitive system. As a solution for bounded rationality I presented fast and frugal heuristics (simple heuristics). I also discussed Optimality Theory, a formal theory that describes the grammar of a language as a set of conflicting constraints that have to be resolved for each formulation and interpretation of utterances. Furthermore I showed a resemblance between the structure of OT and the system of fast and frugal heuristics. I also showed a resemblance between bi-directional OT and some principles from game theory. Taking the viewpoint of game theory, a conversation can be seen as an interpretation game, where both players have interchanging roles of speaker and listener. Furthermore I explored some theoretical concepts in this chapter that I need for my research, such as referring expressions and common ground. Finally I gave an introduction to ACT-R as a modeling environment for a user model. In the next chapter I will discuss the scientific goals of this project.

2.7. Chapter Summary

3. Scientific Goals

3.1. Introduction

So far, I have discussed the theoretical background needed for our research. In contrast to the computational view on cognition, I have explored the theory of bounded rationality. As a solution concept I discussed the use of fast and frugal heuristic rules. These simple heuristic rules can explain complex behavior in decision-making without the use of vast memory and computation. I also explored Optimality Theory and have shown that there are similarities between the hierarchical system of (conflicting) constraints in OT and the system of ordered heuristics in fast and frugal heuristic strategies. Furthermore I discussed referring expressions, common ground and the use of ACT-R for my research.

In this chapter I will discuss the scientific goals of my thesis and formulate my research question. I will also formulate a hypothesis for the research I that wish to perform. My research will concentrate on nominal referring expressions. The experiment setup will be discussed in Chapter 4. I will also discuss the scientific relevance of this project for Artificial Intelligence.

3.2. Research Question

3.2.1. Underlying research question

The concept of bounded rationality can be of influence in cognitive science. Instead of making 'perfect' computational models, we should investigate the boundaries of the human cognitive system and try to explain how these boundaries aid in the formation of our cognitive capabilities. The underlying research question of the project is as follows:

What is the role of bounded rationality in communication?

I will take the analogy between simple heuristics, game theory and OT as a starting point for my research. Communication between humans can be seen as an interpretation-game, in which the players search for the optimal strategy to produce and interpret utterances. OT can be seen as the collection of implicit rules that describe this game. I use bi-directional OT, because this combines both the speaker and hearer point of view. By combining these two roles of a player in a communication-game, bi-directional OT explains the link between syntax and semantics/pragmatics. The overall hypothesis of this project is that the limitations on cognitive capacities play a major role in finding the optimal strategy. The Gricean framework that has been incorporated within bi-directional OT describes this strategy. The basic concept is to avoid the use of too many words, without taking the risk of losing the desired meaning.

My hypothesis is that bounded rationality actually helps in finding an optimal and economic strategy. Because humans have certain limitations on their cognitive system, they are forced to cope with these limitations and to develop better strategies in using the available resources. This way, a cognitive model based on bounded rationality with simple heuristics can actually have an advantage over models that have perfect information, limitless computational power and lots of time to make decisions. (For examples of cognitive models based on bounded rationality, see Todd and Gigerenzer, 1999.)

3.2.2. Hypothesis

To investigate the underlying research question stated above, I will formulate a more specific hypothesis for my experiment. The research will concentrate on the role of bounded rationality in communication with a focus on nominal referring expressions. In pursuing the underlying research question I formulate the following hypothesis:

Linguistic economy plays an important role in selecting a nominal referring expression, when presented with several alternatives in communication.

Often there are several possibilities to refer to the same object or entity in a conversation. These are the different nominal referring expressions. What we want to know is how people select the referring expression of their choice. Why do they choose that specific option and not one of the other possible expressions? In most artificially intelligent models subject will behave perfectly logically and will infer that it is common knowledge that there is e.g. only one white circle at the exact moment this is logically possible. In real life however, people do not make such perfect logical decisions. I expect that subjects will at some point decide (not infer) that something is common ground. Once subjects decide that a certain entity is common ground, they can refer to this subject using a more economical referring expression. This might well be before or after it is logically possible. I think that there are several possible factors that have to be considered here (there might, of course, be more):

- Specificity
- Linguistic economy
- Salience
- Common ground

The principle of specificity demands from the participants that they refer in such a way to a certain object, that there is only one possible candidate in the context as a referent. For example: when there are several red circles, of which only one is large, saying *the large red circle* satisfies the principle of specificity. Other sentences, such as *the red circle* or *the large circle* will not enable your conversation partner to distinguish the object you are trying to bring under his attention (except when one circle is already salient). It is similar to the interpretation constraint BESTRONG from (Blutner, 2000) and STRENGTH from (Zeevat, 2000). There is a direct link between this principle and the common ground of the participants, since the possible candidates of reference are the entities in the common ground.

As described in the theoretical background, participants in a conversation act conform the principle of linguistic economy: both speaker and hearer try to maximize mutual understanding and to minimize efforts. In our experiment I will define the efforts in term of the number of words used by the speaker to describe a certain object. In the section methods I will describe this in more detail. The basic idea of our hypothesis is that because of their bounded rationality, the participants will use fewer words to describe a certain object when possible. This translates to the constraint ECONOMY.

A third factor might be salience. This refers to the availability of items in the common ground as a referent. If two persons are talking about their colleague, and one would say: *she has such a nice suit*, the word *she* is easily interpreted as referring to their mutual colleague, since they were just discussing her: their mutual colleague is a high salience item in the discourse. This is collected in the constraint SALIENCE. I have added this constraint in order to be able to make an analysis of the experiment conversations. This analysis can be found in Chapter 6. SALIENCE is dependent on another constraint: DOAP (Hendriks & de Hoop, 2001) which tells us to seize every opportunity to anaphorize text. First DOAP will try to anaphorize text, then SALIENCE will find the most salient item in the discourse environment as a referent for the anaphoric expression. Here we see a collection of the constraints that reflect the above mentioned factors:

BESTRONG	prefers informationally stronger readings of a sentence.
ECONOMY	prefers simple interpretations.
SALIENCE	prefers the most salient item in the common ground as a referent for an anaphoric expression.
DOAP	Don't Overlook Anaphoric Possibilities.

It is important to notice that all these factors are in some way related. They are interacting with eachother. They all have to be seen in the light of the common ground that has been established at the point of analysis. When a subject is common ground, it then becomes possible to refer to that subject using an anaphoric expression. When it is not, it has to be introduced to the common ground using a more descriptive expression. This will usually require more words than an anaphoric expression, and it will require the hearer to put more effort into his interpretation. This will violate ECONOMY, but it will satisfy SALIENCE, because it introduces a high salience item to the common ground that is suitable as a referent for the used anaphoric expression. Also a subject might become less salient once it has been introduced to the common ground. When a subject has been discussed at the beginning of a conversation and is not mentioned anymore, it will slowly become less salient. This is because of the limited attention span of the human cognitive system. At a later stadium of the conversation it might thus be necessary to re-introduce the subject again to avoid misinterpretation. Even though this would require more words it is an example of linguistic economy: using too few words to describe the less salient subject might end in misunderstanding! So, when one factor changes, it may have impact on all factors.

For this experiment I will restrict the domain of discourse by creating a finite and controllable context for the conversation: the graph completion task. I will discuss this in more detail in the section methods. By controlling the context I limit the factors that influence the choice of the nominal referring expression. Specificity and context are controllable, so I will be able to see the effects of bounded rationality when participants try to find the most economical expressions.

3.3. Scientific Relevance for A.I.

The contribution that this project makes is that it combines the methods and experience of several fields of research. On the one hand there is the field of linguistics with Optimality Theory (and its analogous model in Game Theory). On the other hand there is the theory of bounded rationality and simple heuristics. I use the field of cognitive modeling of learning models (ACT-R) in my experiment. In combining these disciplines we can contribute to AI by investigating the role of bounded rationality in communication. The key objective is to test if humans are restricted or aided by their bounded rationality in finding an optimal strategy. In my experiments I will be able to see these effects

Furthermore, our understanding of the workings of nominal referring expressions will grow. At the current time, I suspect that certain factors play a role in selecting the 'optimal' nominal referring expression, but the exact roles of these factors are not known.

This project is also a pilot study for the research program 'Bounded Rationality in Cognitive Strategies'. This research program aims to propose an integrative theory of human cognition by combining several cross-disciplinary scientific approaches, and by applying them in different areas of cognition. This integrative theory could be based on the theoretical framework I have discussed so far. It can be used to model a cognitive system based on ordered and sometimes conflicting rules. We have seen some evidence for such a system in food choice (the Norway rats using the recognition and illness heuristic rules) and depth perception (Petter's illustion using the overlap and coherence visual cues). This project can aid the research program in finding evidence for a cognitive system based on simple rules. Also it can aid in the development of an actual cognitive model that is able to learn in dynamic and complex tasks. Such a model should be based on simple rules, instead of a complex computational architecture.

3.4. Chapter Summary

In this chapter I have discussed the scientific goals of this project. The overall hypothesis of my research is that bounded rationality (in the form of linguistic economy) plays a major role in communication. My research will focus on the production and interpretation of nominal referring expressions. The basic idea is that participants in a conversation try to avoid the use of too many words, without taking the risk of losing the desired meaning. Furthermore I identified several factors that influence the choice of linguistic items in communication. In the next chapter I will discuss the methods for my research.

4. Methods

4.1 Introduction

In this chapter I will consider the methods I used for my research. I will describe the experimental setup and the model I used. I will also describe the main experiment and I will give predictions for the outcome of this experiment. First of all I will describe the restrictions for my experiment. These restrictions have to be met, in order to find a suitable experiment. Then I will describe the graph completion task. Matessa has developed an experiment using this task. I will show that this experiment meets my restrictions and is suitable for my needs.

Furthermore, I will describe a short testcase. In this testcase two participants solved one puzzle in the graph completion task. I used this testcase to get some general ideas of the puzzle solving process, such as time estimates and language use.

In Section 4.3 I will describe the program as participants used it. Using screenshots of the interface I will describe how participants could interact with the interface.

Even though a model and experiment software have been provided by Matessa, I had to make several adjustments to the program, model and experimental setting. I will describe these changes in Section 4.5.

Then I will describe my experimental settings: an ACT-R model and a participant had to solve several puzzles cooperatively. Also, I will describe in detail what information each participant was given at the start of the experiments.

The final section will conclude with predictions of the outcome of the experiments. These predictions are based on the theoretical background and the methods I used. I will return to these predictions in chapter 6, in which I will discuss my predictions with respect to the outcome of the experiments.

4.2 Finding a suitable experiment

4.2.1. Restrictions for the experiment

In this project I wish to look at the generation and interpretation of nominal referring expressions. Such expressions are generally used in everyday conversation. Ordinarily there are several different nominal expressions suitable to refer to a certain object in the discourse (see Section 2.4). I wish to investigate how humans determine which of the possible expressions they will use, or to which object a certain expression is supposed to refer.

I want the use of language to be natural in my experiments, as in everyday life. Therefore I need an experiment in which not language itself is the goal, but in which language is a means to reach a certain goal. I need a task in which language plays a crucial role.

It is also important that both participants have the same abilities to communicate but different knowledge to share. If this is the case, different knowledge has to become part of the common ground in order to solve the puzzle. In such an experiment participants will have to make references to different subjects in the discourse environment. This discourse environment will have to be a limited environment, which is controllable. A final restriction is that the communication will have to be by using text messages. I wish to use an ACT-R model for communication and this type of communication is more conducive to modeling. Also, by using text messaging, I restrict participants in their language behavior. They will only be able to use text to communicate, instead of using all kinds of extra-linguistic possibilities, such as gestures, word stress etc.

Here is a brief summary of the above-mentioned restrictions for my experiment:

1. Language as a means
2. Same communication abilities
3. Different knowledge to share
4. Make use of definite reference
5. Limited discourse environment under my control
6. Communication via text messages

4.2.2. Graph completion task

The graph completion task as set up by Matessa (Matessa, 2000a; 2000b) is an experiment that meets the above-mentioned restrictions. In this task, participants have to communicate with each other, in order to solve a puzzle. The task itself is to solve the puzzle and participants will have language as a means to complete this task. The basic setup of this experiment thus satisfies my first restriction. Also, participants need to refer uniquely to different subjects (circles) in the task environment. This satisfies restriction 4.

In the task both participants have access to different graphs as shown in Figure 4.1. These graphs are both incomplete parts of the total graph consisting of different colored circles connected by lines. This graph is somewhat similar to the graphs Levelt used to study communicative reference (Levelt, 1982). Since both participants have different graphs that need to be shared, this aspect satisfies restriction 3.

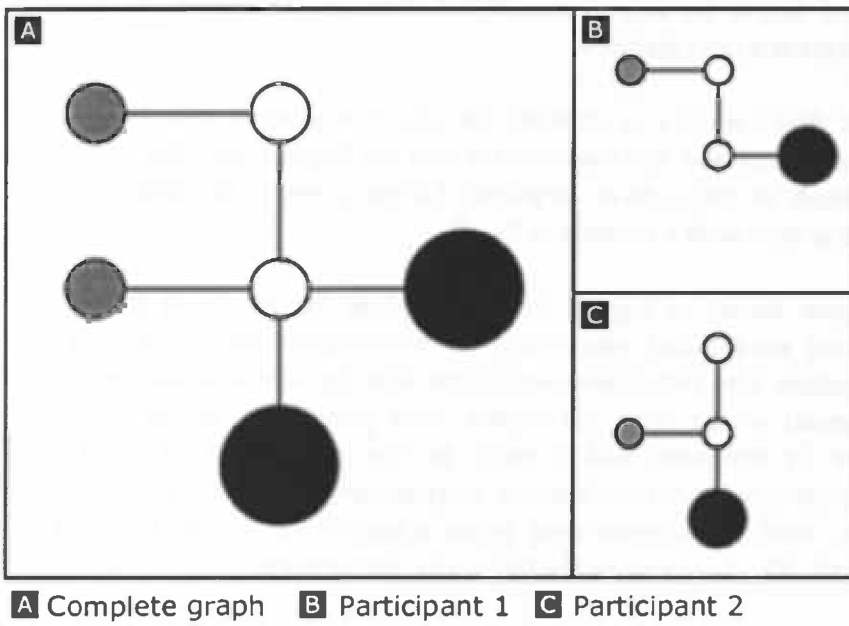

Figure 4.1 Graph example

Participants are each presented with different graphs (either figure B or C). These graphs are overlapping parts of the complete graph, shown in A. In order to solve the puzzle, the participants have to combine the knowledge they each have about their own partial graph.

For communication a computer interface is available and participants can send simple text messages to each other. Participants are restricted in what they can send as text messages to make it possible for a model to respond in a more natural way to the questions or statements of a participant. In game theoretical terms: participants have a finite action set. According to Matessa, this restricted interface need not drastically hinder the communication process (Matessa, 2000a). These aspects satisfy both restrictions 2 as well as 6.

4.3 Testcase

Since I wanted to see some of the language effects at stake, I wanted to create a short test. This test would provide us with some general information about the process of solving the puzzle and possible problems with the puzzle. Also it would give us a good time estimate for solving one puzzle. This was important, because I had to be able to predict the total time necessary for the complete experiment at a later stadium.

Also, at first I wanted to translate the whole experiment into Dutch, since the provided model and software were based on English and the participants all have Dutch as their native language. To see some of the effects in Dutch, I decided to start with a testcase in Dutch.

The figure shown in Figure 4.1 was used as the puzzle to be solved. Two volunteers were found who would try to complete this puzzle via an email conversation. One participant would start with the first sentence by sending this in an email to the other participant. This participant would respond to the question or statement with a reply to this email, and so on. I used this experiment setup, because Matessa's experimental software was still mostly in English. Both participants were given a part of the graph, either B or C in Figure 4.1. They were given the following instructions:

> **Goal:** the goal is to complete the graph. Each of the participants has one part of the graph, consisting of circles with connecting lines. It is sure that one or more of the circles are overlapping the two graphs given to the participants. By communicating with the other, the participants have to complete the graph.
>
> **Restrictions:** participants can only ask or tell about:
> *Connections:* for example "what is above the red circle?"
> *The number of (specific) circles*: for example "I have two (red) circles"
>
> **Confirming:** confirmation of the different circles is done when both participants agree that they have completed the graph. Confirmation can be done using a matrix notation.

Participants can either respond to the question or statement composed by the other participant, or they can formulate their own question or statement. The general idea is that participants may never give information about more than two connected circles. This is to restrict participants from describing their figure completely at once. The instructions on confirming were somewhat more complicated, but I will leave this to the interested reader. The instructions can be found in Appendix 1: "Instructions for the testcase" (in Dutch).

The resulting conversation can be found in Appendix 2: "Testcase email conversation" (in Dutch). The participants succeeded in completing the puzzle, but it took them about 70 minutes to finish, which was longer than expected. The confirmation phase was therefore largely skipped. The extensive timespan used for the completion of the puzzle can be explained by two factors. First, the sending and receiving of email messages was quite slow. Second, the puzzle appeared to be too difficult for the experiment. The last factor was an important one, since I downgraded the overall difficulty of the puzzles in my main experiment for this reason.

4.4 Program description

I will describe the interface of the experiment software here using screenshots of the interface in action. The experiment and interface are also explained in Appendix 6: "Uitleg Experiment" (in Dutch).

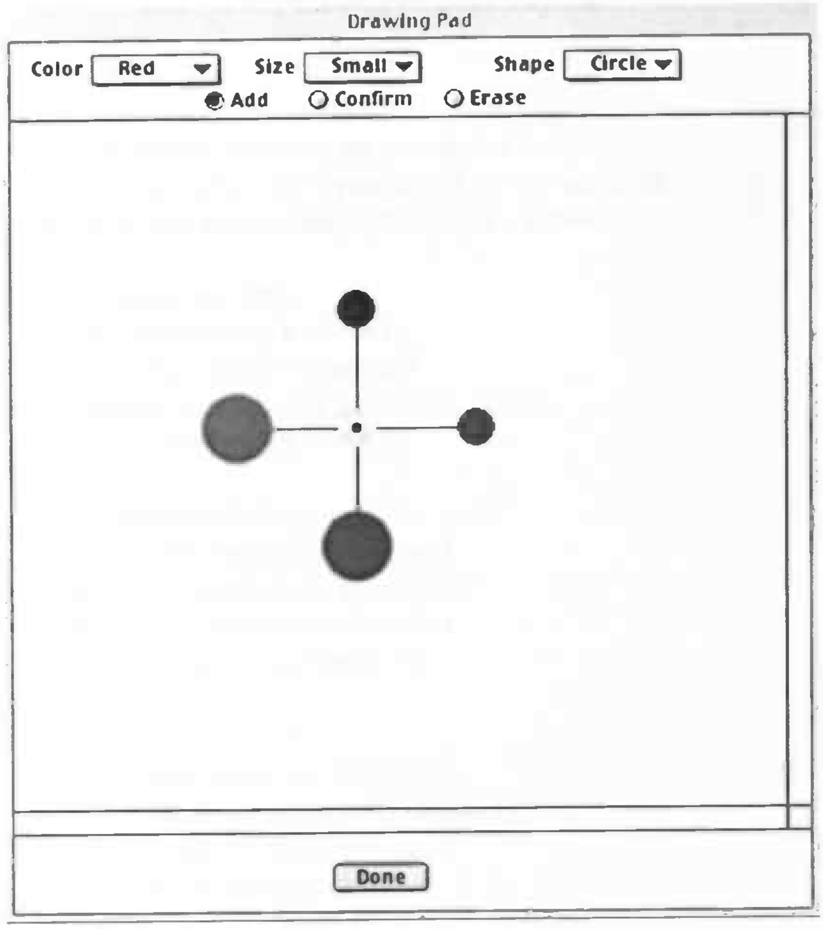

Figure 4.2 The Drawing Pad

In this screen participants can see their figure. Initially this window will contain the given part of the figure. Participants can add new circles when they believe there is another circle in the complete graph. When participants are sure they have completed the graph they can start confirming. During confirmation, participants can select Confirm and click on the circle they want to confirm with their partner. If their partner selects the same circle, the common score will be increased onscreen with 10. If the two circles differ, the score will be decreased with 10. The button 'Done' on the bottom of the screen can be used if the puzzle is completed, i.e. when all circles have been confirmed. If the button is pressed, the next puzzle will be loaded.

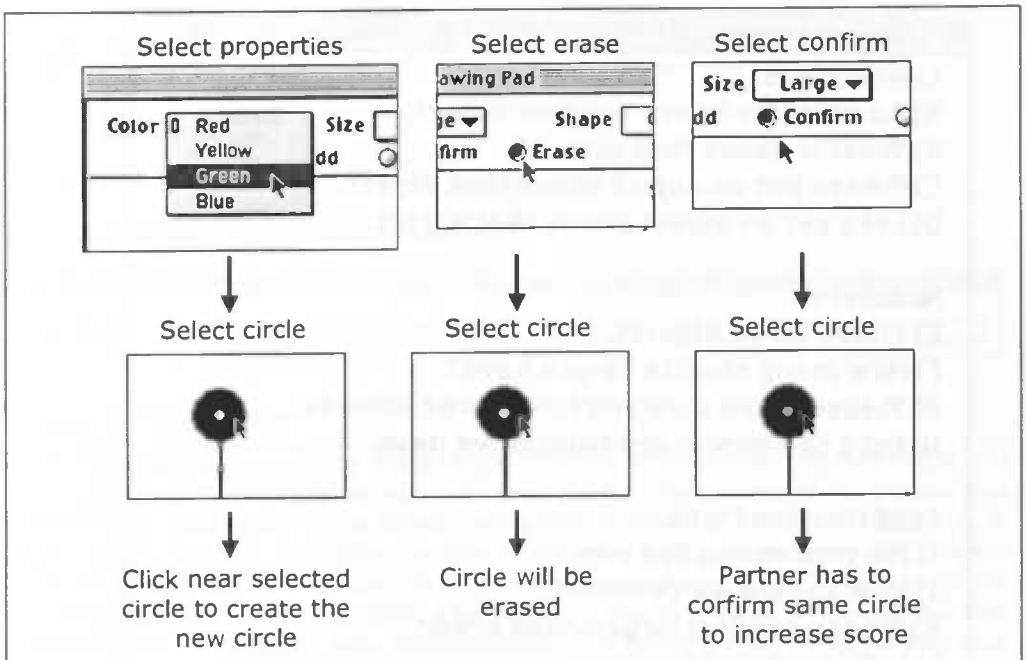

Figure 4.3 Adding, erasing and confirming circles

Adding a circle is done by selecting the right properties (color, size, etc) and selecting the circle to which the new circle must be connected. To place the new circle, participants have to click near and in the desired direction from the selected circle. Erasing and confirming is done by selecting the right option and clicking the desired circle to erase.

Figure 4.4 Score window

This window displays the score of the participants. When participants confirm the same circle, their common score will be increased on the screen immediately. If a circle is selected for communication, the window will display the text 'waiting for you...' or 'waiting for partner...'

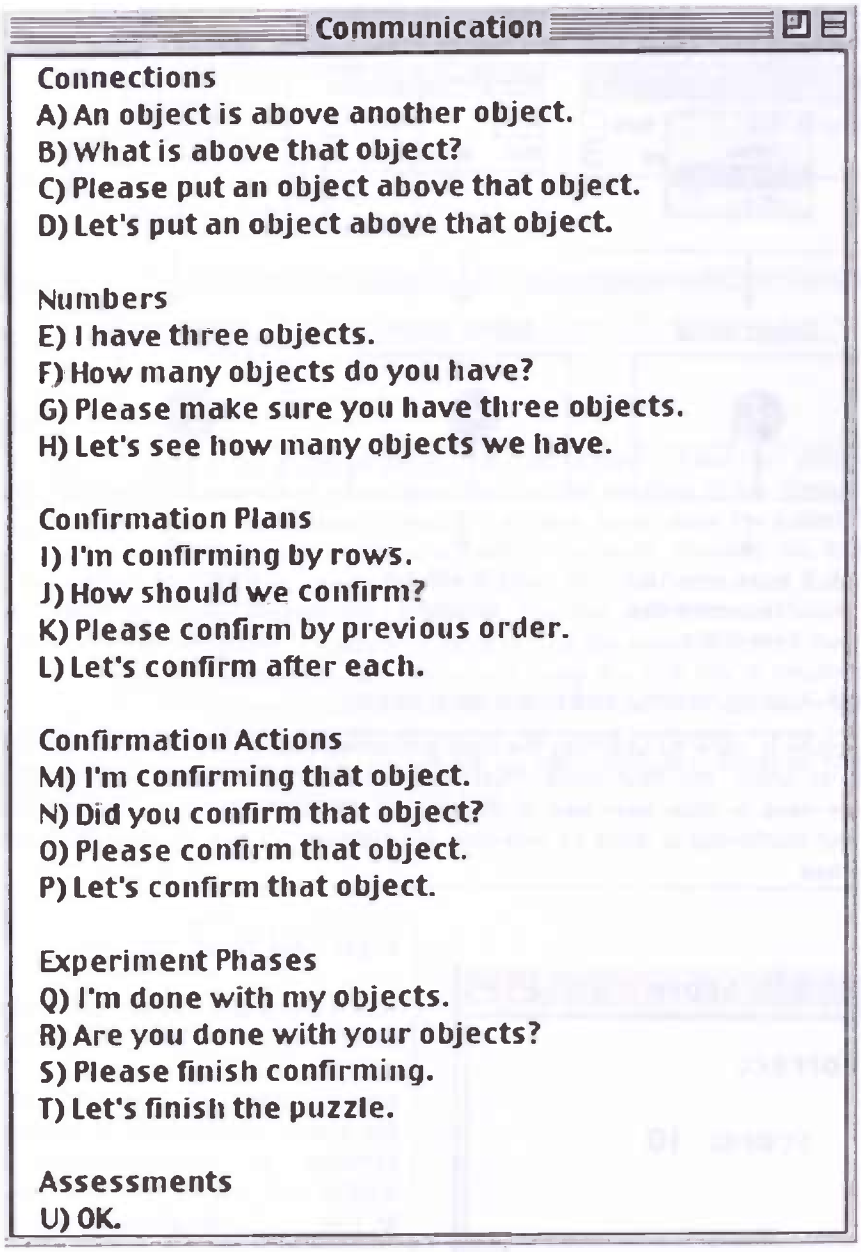

Figure 4.5 Communication window

The communication window shows the types of sentences participants can use to communicate with their partner. The example sentences are grouped by type and can be selected by typing the according letter at the beginning of the sentence. E.g. to select the first type, participants have to type 'a'. The chat window will pop-up and will display the different possible words that can be used to formulate one of the sentences from the range [A-U].

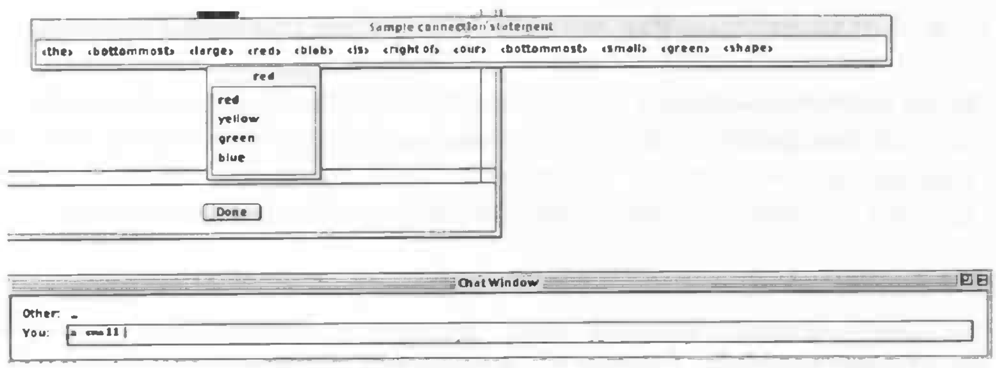

Figure 4.6 Chat window

When participants press a key in the range [A-U] a chat window will become visible and the sample connection window will guide them through the creation of the chosen type of message. Each word in the sample connection window symbolizes the possibility to type such a word. A drop down window is shown at the word, showing the different options the participants can type. When a word is typed the chat window will 'jump' one character ahead by adding a space after the word. The drop down window for the next word is shown at the same time. Participants have to type the entire sentence they wish to send to their partner. Participants can skip words using the TAB-key or return to the communication window (above) by using the ESC-key. The BACKSPACE-key works as usual and pressing RETURN or ENTER will submit the typed sentence. Using the TAB key to skip a word will create a time advantage for the participant, because typing a word will take more time than pressing the TAB key.

4.5 Changes on the model

In the following section, I will discuss the changes I made to the model. A detailed description of the ACT-R model and instructions how to start the program can be found in Appendix 3: "ACT-R model" (in Dutch). This appendix also describes some of the encountered problems with the software.

4.5.1. Translation into Dutch

As mentioned earlier, I wanted to translate the complete experiment and model into Dutch. Eventually it turned out to be quite a problem to do this. The main problem in translating the model was that it is based on a rule-based response mechanism, instead of a grammar. I will set out the obstacles I encountered, that posed such a problem that I decided not to translate the model and the software.

Sentence recombinations

First, the model and interface were based on sentence recombinations, as displayed in the communication window (see Section 4.4). These sentences were in fact recombinations of two basic types of sentences. Basically, the recombinations of sentences were used as the basis for a response mechanism, instead of creating a grammar that the model could use to formulate its response. The response mechanism was created to have the model respond in a realistic way to different types of given sentences. The sentences were therefore divided in 5 semantic types as reflected by the grouping in the communication window:

Paired	about two connected circles (a pair)
Number	about the number of (specific) circles
Conf-plan	about the way the confirmation should take place
Conf	confirmation sentences
Phase	about the phase of the experiment (for ending the puzzle)

These 5 semantic types of sentences were combined with 5 pragmatic types of sentences:

State	a sentence in which participants can state a proposition
Query	a question
Req	a request to the other participant to check or to do something
Prop	a proposal to the other participant to check or to do something
Assess	assessments, like "Ok" or "Yes".

All combinations with assess were the same, so there were in fact 21 different overall sentence types (not 25). Each of the 21 sentence types has a fixed structure with slots for the different words. Most of the words were variable: participants were able to choose from a set of possible words for each slot. The set of words available for one slot generally consists of synonyms or alternatives for a specific word, e.g. *circle, blob, object* or *above, below, right of, left of.* This structure is the basis for the entire model. When a participant creates a sentence, the model is able to create a response message using these type-combinations of the created sentence. So the model uses a response mechanism, instead of a real grammar. This is an important difference, because a grammar could presumably have been changed into Dutch grammar, but this mechanism couldn't be translated. For example, a participant creates a paired-state message, which is a statement about two connected circles, such as:

a large blue circle is above our small yellow circle.

The model will extract the information in the sentence (about the large blue circle) and will detect the two types that were used to form the original message. (This is actually done during the creation of the message by the participant.) Then it will respond to the sentence with – in this case – another paired-state sentence. So the model could for instance respond with:

a small red circle is left of our small yellow circle.

For each combination of types there is a specific response that the model will use.

In such a fashion the way of responding of the model was constructed. The recombination of sentence types is a recognition mechanism for the model to respond to messages. This results in a realistic way of responding to different types of sentences. However, this construction of the model does not allow easy modification of the types of sentences, since the entire model is based on this recombination, instead on a specific grammar for this small domain. To be able to use the model, all recombinations of types would have to be translated into Dutch. However, several of these sentences would be considered unnatural or unnecessary in Dutch.

Other translation problems

In translating words and sentences into Dutch I encountered several difficulties. Translating was much more ambiguous than expected. Examples can be found in Appendix 4: "Other translation problems".

Also, some of the ACT-R production rules were specifically written for certain English sentences. Most of the production rules were in fact quite general, filling in the specific words dependent on the content of the received message. However, somehow specific ACT-R rules were considered necessary to complete the model. These specific production rules contained the actual English words that would have to appear in a message in order for the rule to fire. For example: a production rule could contain the condition that the previous word had to be *how*. Otherwise the rule would fail.

At first it seemed that simply translating these words into Dutch would solve the problem. But, as mentioned above, there are some differences between English and Dutch. This would sometimes make it impossible to simply translate the words, for example when word order in Dutch is different from word order in English. The model is dependent on task-specific ACT-R production rules that have been written for Matessa's experiment specifically.

So, after many hours of coding and testing I decided not to translate the software into Dutch. There was still much to do, as we will see in the following paragraph about redundancy of dimensions. The problems were in fact of such a nature, that it is strongly advised to create a new model for Dutch. I will give these and more recommendations for the future in Chapter 6.

4.5.2. Redundancy of dimensions

In Matessa's experiment objects had redundant dimensions, so that several features were persistent for each object, e.g. red objects were always small and thin. This way there was more redundant information in the problem that could later be left out of the messages. This was an important feature of his experiment, on which most of the ACT-R model's production rules were based. This was a serious obstacle for my experimental settings since I needed non-redundant dimensions to be able to create more ambiguous graphs.

So, this form of redundancy is not very suitable for my experiment. Since I want to investigate the role of referring expressions, the setting of the figures had to be in such a way that such referring expressions can and will be used. For example: when presented with part B of in Figure 4.1, subjects do not know the complete graph. Referring to the rightmost large black circle as "the black one" is not possible, because the other participant (hearer) might have more than one black circle in her partial graph. Referring in such a fashion is only possible when a subject knows or has decided that it is common ground that there is only one black circle.

I had to construct several graphs with non-redundant dimensions. This way, participants are forced to reason about common ground in order to be able to use referring expressions. This allows us to spot the effects of the different factors at investigation here, especially linguistic economy.

In the old model the vast majority of the ACT-R production rules were based on the fact that dimensions were redundant. The old experiment had four dimensions: location, color, shape and size. Each of the last three dimensions was linked to the other two dimensions in the following way: red was always small and thin, green was always medium and round and blue was always large and fat. The location dimension contained information about the relative position of a circle. For example: a circle could be the leftmost or topmost circle. (See the program description in Section 4.4 and conversations for more detailed information.) The location dimension was not linked to the other dimensions for redundancy. I discarded the shape dimension, since I wanted strictly round circles to avoid misunderstandings (ovals are also circles!). So, I needed only the color and size dimensions and added the color *yellow* for more diversity.

Since in the old model the dimensions were coupled in the above-described way, the ACT-R model could easily find a specific circle based upon information in one dimension. If a circle was red, it would also be small and thin, so these aspects of a circle could be filled in by the model. This is exactly how the model worked. So when I created figures with non-redundant dimensions the model wasn't able anymore to find the right referent for a certain expression.

In order to expand the model so that it would be able to find the right referent, I had to create new production rules. These rules had to be able to take all the information available about the referent. If for example, both color and size are available, the model should be able to use this information and look for a circle in the figure that matches these aspects. If on the other hand only color and location were available, the model would have to use this information to select the right circle if possible. I added the desired rules to the software and altered the existing production-rules. Most of these added or altered production rules can be found in Appendix 5 "ACT-R regels" (with Dutch comments).

4.6 Description of the main experiment

In the main experiment 10 participants had to complete 6 different puzzles (see Figure 4.8 on the next page). The experiments took place at the DWAT lab (the Digital Workspace lab at the faculty of Psychology of the University of Groningen). The lab has a control room, with a room for experiments on each side of this room. Both experiment rooms can be observed from within the control room. Participants in the experiment rooms cannot see into the control room, because of the see-through mirrors that have been placed between the experiment rooms and the control room. The setup is illustrated in Figure 4.7.

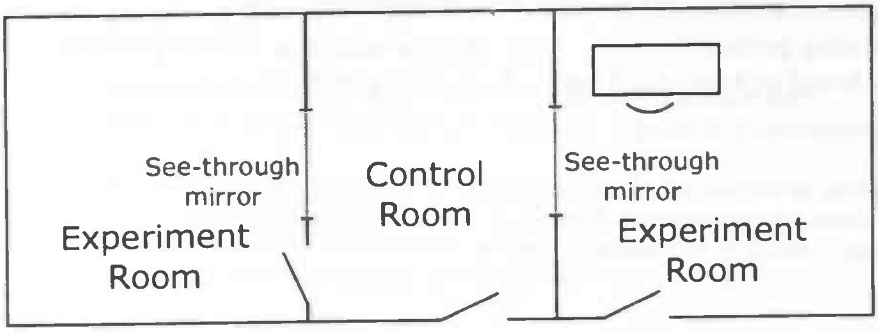

Figure 4.7 DWAT

At the Digital Workspace both experiment rooms can be monitored from the control room via computer screens and the see-through mirrors.

In my setting however, I used only one of the experiment rooms, since the participants were in fact going to interact with my ACT-R model. The setup at the DWAT allowed us to give the participant the impression that they would be interacting with another participant. If participants would have known that they were interacting with a computer model, they would probably fit their communicative behavior to the model. Since I wanted to investigate communication in a natural environment, this would have been an undesired effect.

The setup allowed us to monitor the participants, so I could intervene if something would go wrong. In this way, I could also keep up the impression that there was another participant.

Each of the participants was given a sheet with the instructions. These instructions can be found in Appendix 6: "Uitleg Experiment" (in Dutch). They could try out the interface of the experiment, because a test run was available. In this test run, participants could get used to the different functions of the interface. They were shown how to create the different messages and how to work with the drawing pad. I will describe the instructions that were given to the participants here in English, since these are vital to the experiment.

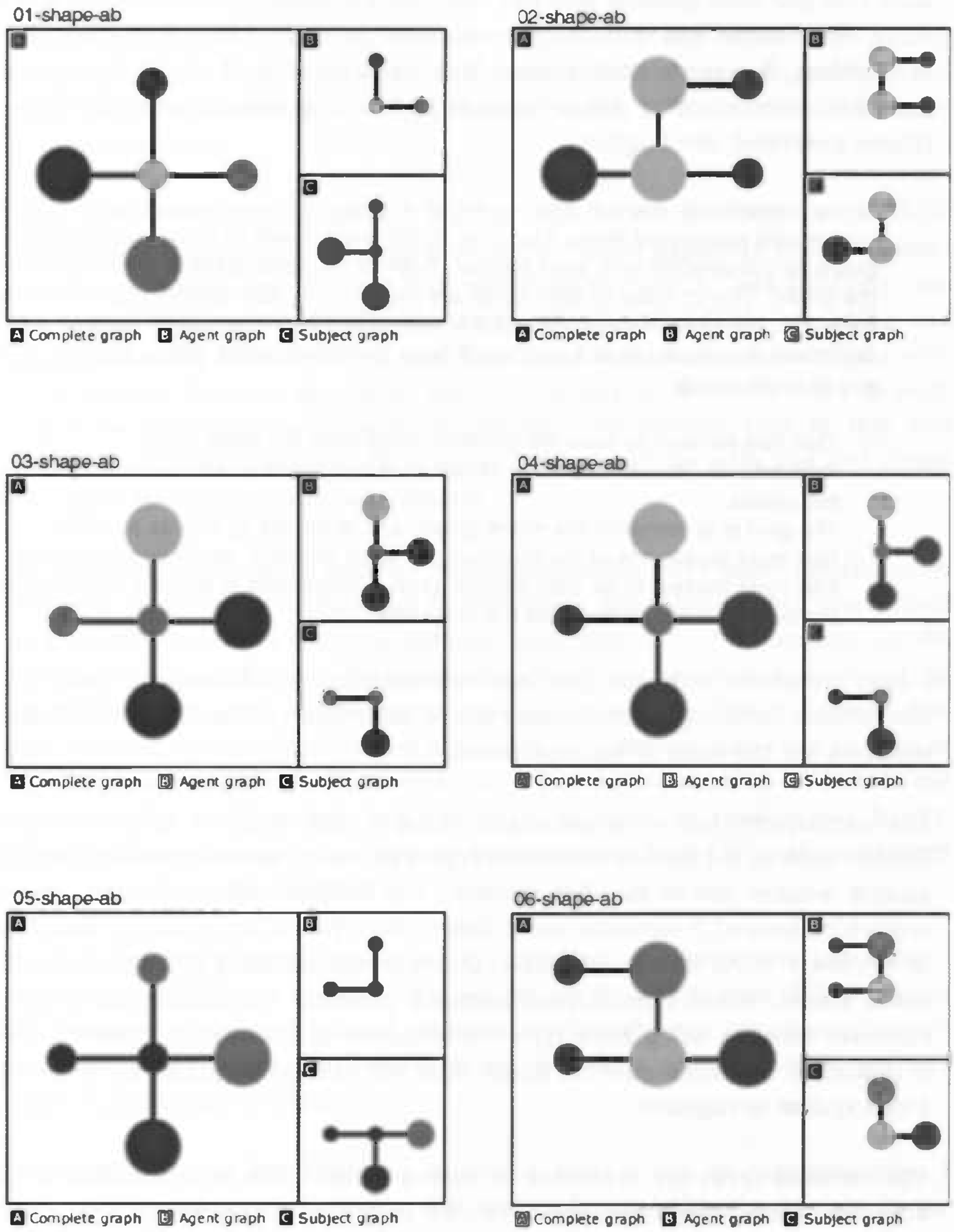

Figure 4.8 The six puzzles for the experiment

As mentioned before, each participant was given a set of instructions. They were told that their partners also had received this set of instructions in order to make the instructions part of the common ground. After they read these instructions, it was checked whether they understood each of the instructions and they were clarified where necessary. The instructions were as follows (freely translated into English):

> In the experiment you will see a part of a graph on your screen, consisting of circles with connected lines. The goal of the experiment is to complete the total graph in cooperation with your partner. Both participants have a different part of the graph. One or more of the circles are the same in both partial graphs. Of one circle you will know in advance that it is common. This circle will be marked with a tiny black dot inside. At the end you'll have to confirm each circle, this will allow you to score points.
>
> - 'Our' can be used to mark the common circle (with the black dot).
> - Relations as 'above' and 'right of' count only if there is a connection between two circles.
> - The goal is to complete the entire graph, and to do this as fast as possible.
> - You must always speak the truth: answer each question with a correct answer.
> - You must always do as your partner says: a player has to place a certain circle somewhere if her partner tells her to do this.

I have translated only the goal and rule instructions above. The sheet with instructions continued after this part with a description of the different interface windows and the order of the experiment.

The participants had to communicate using a chat window. In a turn-based fashion each of the participants could type a sentence, selecting words from the sample window above the chat window. The different sentences were in fact recombinations of 5 semantic and 5 pragmatic types of sentences as described in Section 4. 5. As mentioned before, this response structure is the basis for the entire model. When a participant creates a sentence, the model will create a response message using these type-combinations of the created sentence. This is important to see, because the model does not have a real grammar, but a rule based system to response.

The sentence types are organized in such a fashion that it is possible to use different words within the same type of sentence. These words were either synonyms or alternatives, e.g. *circle, blob, object* or *above, below, right of, left of*. Participants could select a sentence type by pressing a letter within the range [A-U]. The chat window and sample window would then pop-up and the participants were guided through the creation of the sentence word by word. It is possible for the participants to skip a word by pressing the TAB-key. In this way participants can save time by skipping the words that they believe aren't necessary. (For a description of the interface, see Section 4.4).

4.7 Experiment predictions

I will discuss my expectations about the outcome of the experiments in this paragraph. First I will give a brief summary of theory, hypothesis and experimental setting so far, in order to see what this all means for the outcome of the experiment.

What have we discussed so far? We looked at some properties of artificial modeling of human cognition and discovered that AI generally tends to model humans as perfect logical machines with infinite memory and infinitely fast reasoning. We contrasted this with some aspects of real-life reasoning and decision making and concluded that the way of modeling in AI conflicts with these aspects. Humans appear to have limitations in their cognitive system, which we described as bounded rationality. The general idea is that this bounded rationality does not limit us, but that it aids us with fast and frugal heuristics for decision-making processes.

As an example of such a decision making process, I have taken the choice of linguistic items in communication. In Optimality Theory (OT), this is described as a conflict resolution between different constraints. These constraints are the building blocks of our grammar and are ordered in a hierarchical fashion. The conflict resolution in OT has some parallels with game theory. Viewed this way, production and interpretation of utterances in conversations can be seen as an interpretation game, in which each player has the same action-set (words and sentences that can be used) and interchanging roles of speaker and hearer. Important constraints at work here are the constraints from OT, such as the ones from Blutner's bi-directional OT.

Furthermore we discussed some aspects of referring expressions and made a distinction in different types of expressions for my experiment. We discussed the concept of common ground as being deliberately and consciously shared information, of which participants know that their communication partners know it, and so on to infinity.

My basic idea was that bounded rationality plays a role in communication. I formed a hypothesis, which says that bounded rationality can be seen in my experiment in the form of linguistic economy:

> Linguistic economy plays an important role in selecting a nominal referring expression, when presented with several alternatives in communication.

The intuitive idea behind this is that when speakers or hearers have to create or solve a reference in an utterance, they will try to use the minimal amount of words, while still being understandable. They will try to minimize their efforts and maximize mutual understanding.

Recall from Section 2.4 that I distinguished three types of referring expressions: INDEFINITE, DEFINITE and ZERO expressions. I also made a second distinction within each of these types, reflecting the number of adjectives (ADJ) that are used in a certain expression. (For a more detailed description see Section 2.4.)

So if I translate my hypothesis into predictions of what the outcome of the experiments will be, I will be able to check my hypothesis against collected data. What I expect is the following:

> If a participant is able to use fewer words to describe a certain object, (while it is still clear what object is the referent) she will do so.

Now I will have to describe the conditions for when it will be possible to use a shorter expression. As I described in Chapter 3, different factors play a role in the selection of a nominal referring expression, when presented with different alternatives. Some of these factors are:

- Specificity
- Linguistic economy
- Salience
- Common ground

Now let's take a look at these factors once more from the perspective of the experiment. The context in which the conversations take place is the environment of the graphs. For the participant the graph is what she knows about the world and this is the information she will have to share with the other participant. The instructions mentioned above will leave the participant believing that she will have to cooperate with a partner. In fact, there is no such partner, so linguistic variation can be attributed to the participant alone.

The instructions (and the fact that these instructions are part of the common ground) will add some information to the common ground. The participant will believe that the circle with the tiny black dot inside is the common circle. The use of the word *our* is restricted to this common circle. Furthermore the participant will have information about the use of the relation-words, such as *above* or *right of*. The participant will believe that the goal is to complete the puzzle, and that she has to cooperate with her partner in order to complete this goal.

Since all this information is in the instructions, and the participant believes that her partner has also read these instructions, the participant will believe that her partner believes the above and that her partner believes that she believes the above, and so on. Thus: the information in the instructions is now part of the common ground. This is important: since I developed the graphs and the instructions, I now know what is part of the common ground at the start of the experiment. As mentioned in the restrictions for my experiment, I now have a limited discourse environment under my control.

The alternatives for reference are constrained by the words and sentences participants can use. In real life, this is theoretically an infinite action set: there are infinite possible utterances that can describe one referent. This interface however, restricts the participants in their language use (see Section 4.4), so that they have a finite action set. If a participant wants to refer to a certain object in the discourse environment (the graph) she can only use certain types of sentences and certain words (the sentence types from the Communication window and the words from the Chat window). This allows us to formulate the possible alternatives for the same reference when necessary.

Now I can formulate my predictions for the experiments. The first prediction concerns the addition of objects to the common ground.

The use of indefinite expressions will decrease within each puzzle, while the use of definite expressions and zero anaphora will increase.	**Prediction 1**

Participants will use indefinite expressions to introduce objects. Once introduced, objects are added to the common ground. Participants can now use definite expressions and zero anaphora to refer to the same object, since the objects are grounded. I expect to see this effect within each puzzle.

The second prediction is as follows:

In more ambiguous puzzles, participants will use more adjectives to describe objects compared to less ambiguous puzzles.	**Prediction 2**

I have constructed a total of 6 puzzles. The 6 puzzles have been created pair-wise in such a fashion that there is a more and a less ambiguous version of each puzzle. More ambiguous puzzles contain at least two connected circles with the same color and size. In less ambiguous puzzles every circle has a different combination of size and color. The distribution was as follows:

Less ambiguous	More ambiguous
Puzzle 1	Puzzle 5
Puzzle 4	Puzzle 3
Puzzle 6	Puzzle 2

I expect that in the more ambiguous puzzles (5,3,2) the participants will need to use more adjectives than in the corresponding less ambiguous puzzles (respectively 1,4,6) in order to be able to distinguish between the different circles. In the less ambiguous puzzles it is possible to be more economical, because it is easier to identify the right object with less adjectives. The puzzles were presented to the participants in a fixed order.

The third prediction is as follows:

Within each puzzle, over time, participants will use fewer adjectives to describe the same referent.	**Prediction 3**

This prediction says that the use of adjectives will decrease when participants are referring to the same object. An object that has been just discussed is highly salient and this makes it possible to refer to this object using fewer adjectives, while still being specific enough.

4.8 Chapter Summary

In this chapter I have described the experimental setup that I used in the experiments. I described the restrictions for a suitable experiment and I described an experiment that could meet these restrictions. I gave a detailed description of the model and the program interface, and showed that it was not possible to translate the software into Dutch, the native language of the participants. Furthermore I described the instructions that were given to each participant and I discussed the effects of these instructions on the experiment participants. Finally, I reflected on the theory and hypothesis that I discussed in earlier chapters and I made predictions for the outcome of the experiments based on the theory, hypothesis and experimental setup. In the next chapter I will discuss the actual outcome of the experiments.

5. Results

5.1 Introduction

In this chapter I will present the results of the experiments I performed. In the previous chapter I discussed the methods and experimental setup I used for the experiments. The results will be presented in the form of graphs with explanatory comments. In Chapter 6 I will discuss these results with respect to the theoretical background set out in Chapter 2, the hypotheses in Chapter 3 and the experiment predictions in Chapter 4.

5.2 Presentation of the results of the experiment

5.2.1. Indefinite versus definite expressions

In this chapter I will display several graphs in which the data from the experiment is presented. There were 10 participants who had to complete 6 puzzles. Some of the participants did not succeed in completing all of the puzzles. The distribution was as follows:

Puzzle	Number of participants
1	10
2	9
3	9
4	9
5	8
6	7

The data consists of the nominal referring expressions used by the participants. Only the expressions of the participants were counted, since the model is not a natural language producer. The expressions were divided according to the distinction made in Paragraph 2.4.4. Zero anaphora were extremely rare, so these are not shown in the results. I will discuss this in Chapter 6.

I have divided each of the conversations into two parts: the discussion stage and the confirmation stage. The discussion stage is when participants discuss their graphs with each other in order to complete the total graph. This is the first part of the conversation. The confirmation stage is when participants have decided that the graph has been completed. This is the second stage of the conversation. At this point the participants will try to establish which exact circle they will confirm in order to increase their score. To this end, they will refer to the circles that have been introduced in the discussion stage of the conversation. If they have succeeded, the graph will be complete in the confirmation stage.

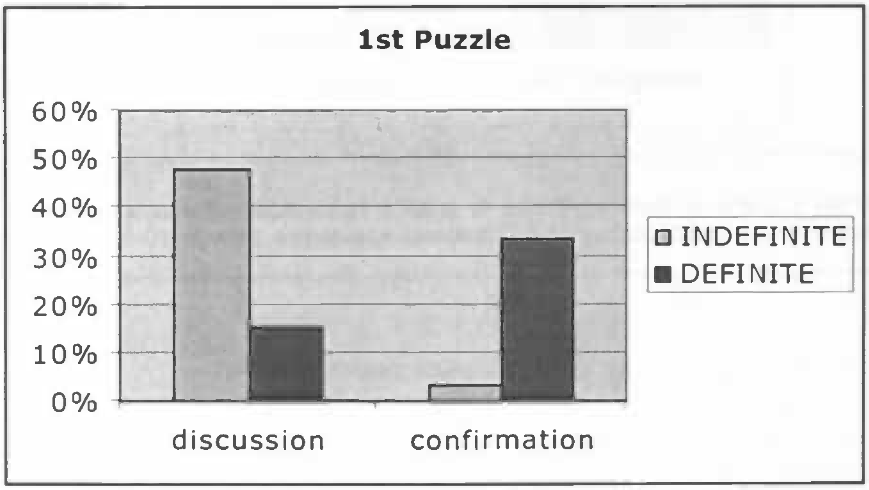

Figure 5.1
This figure shows that the number of definite and indefinite expressions in the first and second stage of the conversation. The figure shows the percentages of definite and indefinite expressions of the first puzzle. Each puzzle showed a similar shaped graph as is shown here: the number of indefinite expressions was always larger than the number of definite expressions in the discussion stage and vice versa in the confirmation stage.

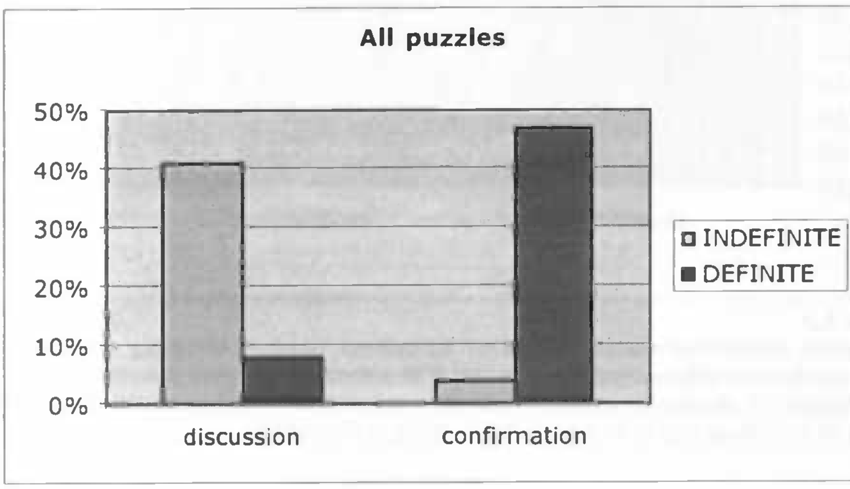

Figure 5.2
This figure shows the percentage of definite and indefinite expressions during the discussion and confirmation stage of all puzzles combined. It shows that the overall results resemble the results in the graph in figure 5.1.

5.2.2. Number of adjectives

From the experiment the number of indefinite expressions, definite expressions and zero expressions were counted. These three types of expressions were subdivided in four other types as is described in Paragraph 2.4.4. These four types were: 3 adj, 2 adj, 1 adj and 0 adj, simply reflecting the number of adjectives in the expressions. To see the difference in adjective use between different puzzles, the two conversation stages mentioned above were combined per puzzle. The expressions were now ordered according to the four adjective types. The number of adjectives was calculated from the number of expressions, multiplied by the number of adjectives in such an expression. E.g. 10 expressions of the 3 adj type result in 10*3 = 30 adjectives.

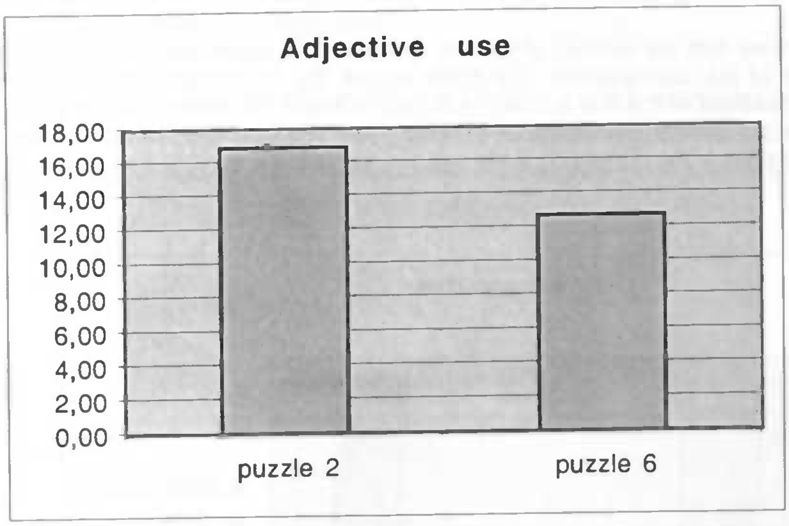

Figure 5.3
This figure shows the mean number of adjectives used in referring expression per participant for two different puzzles. Puzzle 2 is a more ambiguous puzzle than puzzle 6 (see Chapter 4). Since not every puzzle was completed by every participant the mean number of adjectives had to be calculated for each of the puzzles.

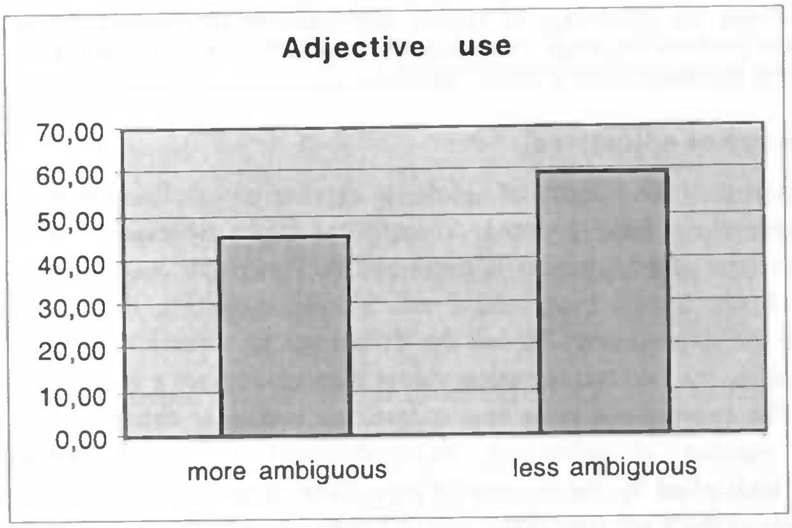

Figure 5.4
This figure shows the total mean number of adjectives used in referring expression per participant for the more ambiguous puzzles (2,3,5) versus the less ambiguous puzzles (1,4,6).

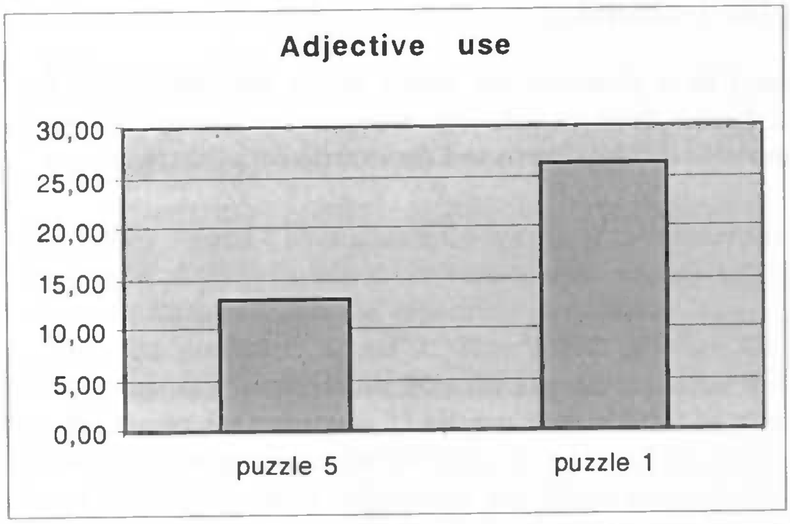

Figure 5.5
This figure shows the mean number of adjectives used in referring expression per participant for two other puzzles. Puzzle 5 is a more ambiguous puzzle than puzzle 1 (see Chapter 4). This was a somewhat more surprising result, as we will see in Chapter 6.

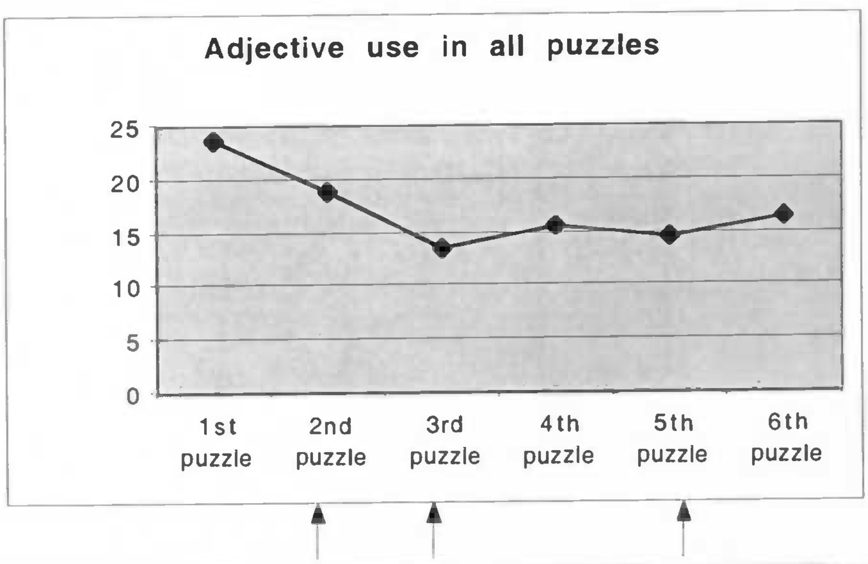

Figure 5.6
This figure shows the mean number of adjectives used in referring expression per participant for all the puzzles. The puzzles are set out in a chronological order. The puzzles marked with the arrows are the more ambiguous puzzles as is shown below.

Less ambiguous	More ambiguous
Puzzle 1	Puzzle 5
Puzzle 4	Puzzle 3
Puzzle 6	Puzzle 2

5.2.3. The conversations

The actual conversations can be found at the following web page http://www.ai.rug.nl/~stinger/afstuderen. This web page shows the complete conversations of each of the experiments.

5.3 Chapter Summary

In this chapter I have presented the results of my experiments. In the next chapter I will discuss these results with respect to the theoretical framework, the research question and hypothesis and the experiment predictions.

6. Discussion

6.1 Introduction

In the previous chapter I have shown the results of the experiments I performed. In this chapter I will discuss these results with respect to the hypotheses I formulated in Chapter 3 and the experiment predictions from Chapter 4. I will first reflect on prediction 1 and 2. Then I will give an OT analysis of a conversation about a puzzle taken from the experiment. After this analysis I will discuss the results for prediction 3. All participants filled in an evaluation form. The answers and suggestions the participants gave will be discussed also in this chapter. Furthermore, some suggestions for future research will be discussed. Finally I will give my conclusions for the research experiment and the results.

6.2 Interpretation of the presented results

6.2.1. Prediction 1

My first prediction concerned the use of indefinite and definite expressions and null anaphora. I expected that the use of indefinite expressions would decrease during one puzzle, while the use of definite expressions and null anaphora would increase. The use of null anaphora was very rare in the experiments, since it was not possible to use syntactically correct null anaphora in the experiment. This was because the types of sentences and words simply did not allow for a syntactically correct English null anaphor. I considered this option because in Dutch null anaphora have a different syntactic structure than in English. (In Section 2.4 I already mentioned this phenomenon.) At the start of this project I wanted to use a Dutch model and a Dutch experiment. In that case the use of null anaphora would have been likely. Eventually I performed the experiments with an English model. Even though the use of null anaphora wasn't syntactically correct in this experiment setup, Dutch participants could have used the Dutch 'form' in the English language. This means that participants would use expressions such as *the blue*, which is syntactically correct in Dutch. In contrast to English, the use of Dutch null anaphora in this setup has a clear economical advantage over the use of Dutch definite expressions. However, the participants hardly used these forms. Either they were aware of the incorrectness of such a form in English, or they simply did not consider the possibility.

This means that I looked only at the number of indefinite versus the number of definite expressions. To see whether there was a change in the use of these expressions, I divided the conversation in a discussion and a confirmation stage as described in Chapter 5. The discussion stage was the first stage of the conversation and the confirmation stage was the second stage of the conversation. If we look at figure 5.1, we see that in the first puzzle my prediction corresponds to the data. In the first stage (discussion) we see a large percentage of indefinite expressions, versus a small percentage of definite expressions. In the second stage of the conversation (confirmation) we see the opposite effect. The other five puzzles showed a similar shape in the distribution of indefinite and definite expressions. This is collected in figure 5.2. It shows that in general the use of indefinite expressions is concentrated in the discussion stage and the use of definite expressions is concentrated in the confirmation stage. The data thus correspond with my first prediction.

6.2.2. Prediction 2

The second prediction I made concerned the use of adjectives. The manner in which the number of adjectives was calculated is described in Chapter 5. In figure 5.3 we see the number of adjectives used in puzzle 2 (more ambiguous) versus puzzle 6 (less ambiguous). The puzzles are shown in Figure 4.8. In this figure the distribution of adjectives is as predicted: more adjectives are used in puzzle 2, which is the most ambiguous variant. But if we look at figure 5.4, we see a different distribution. In this figure the total mean number of adjectives of the more ambiguous puzzles is compared to the total mean number of less ambiguous puzzles. It is clear that the number of adjectives is larger in the less ambiguous puzzles. This does not correspond to the prediction I made.

The main explanation that figure 5.4 does not correspond to prediction 2, I believe can be found in figure 5.5. We see here that puzzle 1 (less ambiguous) has a drastically higher mean number of adjectives than puzzle 5 (more ambiguous). Why is this? The conversation data (see the web page http://www.ai.rug.nl/~stinger/afstuderen/) indicate that the conversation about the first puzzle was significantly longer than the other conversations. We see a slightly similar effect with respect to the second puzzle. This effect can also be seen in figure 5.6. This figure shows the mean number of adjectives used per participant to complete a puzzle. The puzzles were presented to the participants in a fixed order, so we can see what happens over time in the mean number of adjectives used by the participants. We see a general decrease over time in the number of adjectives used by the participants. I believe this indicates strategy learning in the participants. This presumably means that the participants need to perform one or more puzzles before they have effectively learned how to solve a graph completion puzzle.

The data I collected cannot make a very strong case about strategy learning, but the presented results might be an interesting lead for further research. I believe that it is not unusual for participants performing a certain task to adopt a strategy for this task. What I have investigated can be seen as a local strategy: I looked at the effects within each puzzle, or comparing two puzzles. But the data I presented indicates there might be a more global strategy for the participants in solving graph completion puzzles.

6.3 Analysis of a conversation

6.3.1. The analysis

I will give a bi-directional OT analysis of the following conversation from one of the experiments using some of the constraints I have described in Chapter 3. It is important to remember that the participant is producing her sentences with the goal of making her partner understand her graph. If she succeeds, they will be able to confirm each of the circles and collect points. In reality her partner is an ACT-R model and will not be able to interpret what she is saying. But since the participant does not know this, she will be producing her sentences as if she were communicating with another human participant. The conversation I will look at concerns the discussion and confirmation of the first puzzle. The conversation was about the puzzle shown in Figure 6.1.

The conversation was as follows:

P: the participant
A: the ACT-R model

1) P: i have four objects
2) A: i have three objects
3) P: a small blue object is above our object
4) A: a small red object is right of our object
5) P: a large green circle is left of our object
6) A: ok
7) P: a large red circle is below our object
8) A: ok
9) P: did you confirm the object
10) A: no
11) P: how many circles do you have
12) A: i count five circles
13) P: i'm done with the circles
14) A: could you confirm the small blue object
15) P: please confirm the green object
16) A: i'm confirming a large green object
17) P: could you confirm the bottommost red circle
18) A: would you confirm our small yellow circle
19) P: did you confirm the rightmost small red circle
20) A: no
21) P: how about we finish our puzzle
22) A: ok

Table 6.1 conversation between ACT-R and participant

The first sentence I will look at is 3). We can see that the speaker has chosen to use an indefinite expression here: *a small blue object*. The reason seems simple: the object that she is referring to (the topmost small blue circle) is not part of the common ground yet. At this point we can easily see what the common ground is. It consists of the following items:

{the common circle, four objects, three objects}

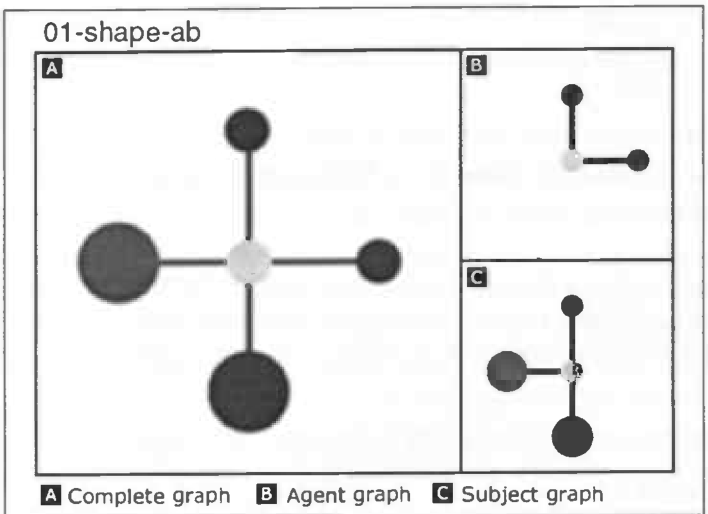

Figure 6.1 the first puzzle of the experiment

The common circle is already part of the common ground at the start of each experiment as a result of the rules of the game that are known to the players. The other two items have been introduced in sentences 1) and 2). For practical convenience I collect these items as *known items*. It is the goal of the speaker to introduce the topmost small blue circle as a new item to the common ground, so that the other player will start believing that this item is part of the complete graph, conform the goal of the experiment. I collect the items that are not part of the common ground yet as *unknown items*. In Section 2.4 I have shown a scale that shows different expressions with respect to the accessibility of a topic. We can use this scale to provide a distinction between known and unknown items:

Known items
↑ Null anaphora
| Definite noun phrases
↓ Indefinite noun phrases
Unknown items

As I mentioned above, participants could not use null anaphora in this setup. Therefore I have left these out of consideration. I will now formulate a constraint that reflects this scale.

GIVEN choose definite expressions above indefinite expressions to refer to known items

I will now explain that this constraint generally blocks the use of a definite expression in the case of an unknown item. The speaker thus uses the expression *a ... object* here, because the referent of the expression is not yet available for the hearer. A definite expression will generally refer to a referent that is available to the hearer. (This is not always the case, e.g. *the queen has her birthday today* is a correct way to introduce *the queen* to the conversation.)

A second constraint at work here is CONSISTENCY (from Zeevat, 2000):

CONSISTENCY do not conflict with the context

In this experiment it entails that one should not conflict with the existing common ground. The tableau in Table 6.1 shows the interaction of the two constraints I have just given.

If the speaker wishes to communicate about a *known item*, both expressions are available, but the use of a definite expression is better according to GIVEN. From the hearer perspective: if a definite expression is being used here, CONSISTENCY will prefer a known item.

			GIVEN	CONSISTENCY		GIVEN	CONSISTENCY
Forms	*Definite expr.*	⇝ ☞					*
	Indefinite expr.		*		⇝ ☞	*	*
			Known items			*Unknown items*	
				Interpretations			

Table 6.1 Tableau for known and unknown items

Definite expressions are coupled with *known items* and this input-output pair will thus block the selection of definite expressions if the speaker wishes to communicate about *unknown items*. If the hearer hears an indefinite expression, she will have to assume that a new object is being introduced, since otherwise the speaker would have used a definite expression. This shows that definite expressions are blocked in the case of an *unknown item* by the optimal input-output pair definite expressions and *known items*.

At the same time, this tableau also accounts for the opposite effect: in sentence 15) we see that the participant refers to a similar object (the leftmost large green circle) with a definite expression: *the green object*. Note that at this point in the conversation, the referent has become a known item, for it has been introduced to the common ground in sentence 5). Possible referents (all known items) for the definite expression in sentence 15) are:

> {the common circle, the topmost small blue circle, the leftmost large green circle, the rightmost small red circle, the bottommost large red circle, four objects, three objects, five circles}

All options except the leftmost large green circle will be eliminated by CONSISTENCY, because every referent (except the leftmost large green circle) will conflict with the current common ground, since the properties of these referents do not match the properties given in the definite expression in sentence 15).

It is important to notice that the constraints are soft. They can be violated by other stronger constraints. The constraint GIVEN I introduced can also be violated. Otherwise it would not be possible to refer to a known item using an indefinite expression.

6.3.2. Prediction 3

My third prediction concerned the number of adjectives used to describe a referent. It says that the use of adjectives will decrease when participants are referring to the same object. In the graph completion task, participants generally need to refer to a certain object in the discussion stage to introduce the object and they need to refer to the same object again in the confirmation stage to confirm the object. We can see this in the conversation above. In sentence 5) (see table 6.1) a leftmost large green circle is introduced with the indefinite expression: *a large green circle*. Two adjectives are used here to introduce the object. From sentence 13) we can infer that the participant has decided already that the graph has been completed. Her graph now resembles the complete graph in figure 6.1. The participant refers again to the same object in sentence 15): *the green object*. One adjective is used here to refer to the object again. This example corresponds to the third prediction I made.

At the moment the participant types sentence 15), the common ground consists of the following objects:

> {the common circle, the topmost small blue circle, the leftmost large green circle, the rightmost small red circle, the bottommost large red circle, four objects, three objects, five circles}

The participant could have used more adjectives to describe the referent (the leftmost large green object), but using only the color is sufficient here because there simply is no other green object. The expression used can be interpreted according to BESTRONG: the listener will interpret the expression *the green object* as strong as possible. This means that the referent will be the leftmost large green circle.

Let's take a look at a different example. In the complete graph there are two red circles: the rightmost small red circle and the bottommost large red circle. In this case we see that the participant does not use fewer adjectives to describe the same referent. In sentence 7) the bottommost large red circle is introduced with the indefinite expression *a large red circle*. Two adjectives are used to describe the object. In sentence 17) the participant refers again to this object, now knowing that there are two different red circles (see the common ground above). The expression she used is: *the bottommost red circle*. Now there is need to be more specific, otherwise the referring expression chosen by the participant might be ambiguous for her partner. For example, if the participant had used the expression *the red circle,* the listener would not be able to identify the correct referent using BESTRONG, because there are two red circles.

Note that if the bottommost large red circle had just been discussed, it would have been highly salient. If the other red circle would not have been highly salient at the same point in the conversation, the participant could now have used a more economic expression such as *the red circle, the circle,* etc. to refer to the bottommost large red circle, according to the constraint SALIENCE (see section 3.1). In sentence 9) we see an example that illustrates the effect of SALIENCE: for the participant it has become possible to refer to an object using only the expression *the object*.

Other examples of participants using fewer adjectives to refer to the same referent are available in the data. These examples correspond to my third prediction about the use of adjectives. Other examples, in which the number of adjectives does not decrease when participants refer to the same referent, suggest that salience of items is an important aspect in choosing the referring expression. If an item is significantly more salient than other items, it is possible to use a more economical expression (e.g. sentence 9). Otherwise it is necessary to use more adjectives to describe the object, because the hearer will not be able to distinguish between the different alternatives for interpretation.

6.4 Evaluation of the experiment

6.4.1. Evaluation questions

Ten subjects in total participated in the experiment. Generally it took them about one and a half hour to complete the six different puzzles, although only seven subjects completed the total of six puzzles. Afterwards they had to fill in an evaluation form of which I will present the results here. To evaluate the experiment I asked the participants to answer the following questions (original questions were in Dutch):

1. Was the explanation sufficient? If not, what was missing?
2. Was the experiment clear to you? If not, what was unclear?
3. Were all the words understandable? If not, which words were not understandable?
4. Did you understand everything your partner said? If not, can you give an example?
5. Could you say everything you wanted to say? If not, what couldn't you say?
6. Was your partner cooperative in solving the puzzle?
7. Is your partner male or female?

6.4.2. Answers to the questions

I will discuss the answers of the participants here briefly.

1

Most participants answered that the forehand explanation of the experiment was sufficient. Two out of ten participants answered that some extra practice in using the interface would have helped them some more. Some of the other participants noted also in the conversation afterwards that completing the first puzzle was more difficult than the other puzzles, because they still had to find out how the interface worked. This corresponds with the idea of global strategy learning I discussed above (prediction 2).

2

Only one participant answered that confirming the objects was difficult during the experiment. Even though the participant had to get used to the interface and the way of solving the puzzle, eventually the experiment was clear.

3

One participant noted that there is a certain discrepancy between the sentences *Let's finish the puzzle* and *Let's finish confirming*. A better distinction might be useful.

4

Some participants answered that they did understand what their partner said, but that they thought that their partner didn't understand them. Other comments included that the models use of the word *our* was incorrect; that the other participant tried to finish the puzzle too early; that the number of objects the model said to have was incorrect and that the models answers were not logical.

5

Participants couldn't say everything simply because the interface restricted the possible words and sentences. Still there were some interesting suggestions from the participants, such as to be able to talk more globally about the process of solving the puzzle; that there is no good reaction possible to the sentence *I don't know*; that there was no possibility to ask where a certain object is in the figure, such as *Where do you see the blue circle?* These might be interesting suggestions for a new puzzle solving task.

6

Six out of ten participants were not satisfied with the cooperation of their partner. Some of the participants were even a bit angry with their 'partner'. Participants said that the initiatives came mostly from their side and that their partner did not always understand what they wrote.

7

Four participants thought their partner was a male; one thought that the partner was a female; three had no clue whatsoever and two thought that the partner was a computer program. Still, when I asked them, most of the participants weren't really convinced that there was a real person on the other side. Apparently the communicative behavior of the model did not convince the participants that their partner was a real human, as is also shown by the answers the participants gave in 4 and 6.

6.5 Suggestions

6.5.1. A new model

In Chapter 4 I discussed several problems I encountered in the model and the software. Since the ACT-R model has been used solely to create an artificial partner for the participants, it cannot be used as a basis for a cognitive agent in communication. The current ACT-R model is based on the recombination of sentence types and uses these recombinations to formulate a response to the sentences that are typed by the participants. To create an actual communicative agent a different approach is necessary.

In my experiments native Dutch speakers had to communicate in English. It would have been better if they were able to communicate in their native language. For this, it would be necessary to create a Dutch artificial partner for the experiment setup I used. I would recommend creating a small grammar specifically for the experiment setup that will be chosen. This grammar should be modeled using simple bi-directional OT constraints as rules. The OT constraints I have discussed are specific for the type of task I used, so more constraints need to be identified in order to create a grammar. If the agent uses such a grammar to 'interpret' and 'produce' its utterances, this will result in a more realistic way of responding than the current model can provide. I believe that if such a communicative agent is based on simple heuristics in the form of constraints, it can account for the choice of linguistic items. This is my first recommendation for a new model.

My second recommendation discusses the creation of a new model taking a wider viewpoint. The research program 'Bounded Rationality in Cognitive Strategies' aims to propose an integrative theory of human cognition. Such an integrative theory can be based on the theoretical framework I have discussed so far: a cognitive system using ordered and sometimes conflicting rules. In order to create an actual cognitive agent in ACT-R based on this theoretical framework, such a model would have to use heuristic rules. A model based on simple rules has a clear advantage over a model based on complex calculations. It will use less computation to achieve a similar result. It would probably not find the optimal result, but it would find a satisfying result, fit for the circumstances. Such a model will make it possible to generate a real-time response.

6.5.2. Global strategy learning

Looking for a global strategy is an interesting subject for further research. A similar task could be used, in which participants have to complete several puzzles over time. To see the effect of learning a global strategy, the puzzles need to be compared in a linear fashion. To see the effects of more ambiguous puzzles versus less ambiguous puzzles, it is necessary to use a different ordering of the puzzles for different participants. Based on my observations I believe that at least one or two puzzles have to be completed in order to adopt a good strategy for this task. Participants simply have to get used to the interface and the game that is played. If the order in which the puzzles is varied, the effect of global strategy learning is not restricted to one or two puzzles. The latter is the case in my experiment, since the order of the puzzles is fixed. Further research could give a good indication of the effects of global strategy learning and the number of tasks that is necessary to adopt a good strategy. One might ask what defines a good strategy. In the graph completion task a good strategy scores the highest number of points in the minimal amount of time.

6.5.3. Towards a single OT model

Blutner uses the kind of tableau I used here, to explain blocking effects. He takes both the speaker and the hearer point of view (Blutner, 2000). His point is that structures that compete in the first point of view are constrained by the outcomes of the other point of view and vice versa. Zeevat gives a similar argument, saying that not only the hearer, but also the speaker harmonize between the interpretation and generation process (Zeevat, 2000). The case made here is quite simple: if the hearer is able to optimize with the use of a speaker point of view, why wouldn't the speaker optimize using the hearer point of view? It isn't smart to understand something of which you know that the speaker does not want you to start thinking it. And it isn't smart to say something, when you know that it will not be understood well. Both Blutner and Zeevat seem to aim at a single optimization model that incorporates both the hearer and speaker point of view simultaneously. I believe that such a model should be the basis of a cognitive agent for communication.

6.6 Chapter Summary

I have now discussed the results from my experiments with respect to the predictions I made in Chapter 4. Using the theoretical framework that I have presented, I have analyzed and explained a conversation from the experiment. The constraints from OT used as simple heuristics make it possible to account for the effects seen in the data. Interesting new questions have arisen, such as what the effect of global strategy learning might be in such an experiment. Furthermore I discussed an evaluation of the experiment by the participants. And I have made some suggestions for further research.

7. Conclusions

I started this thesis with the remark that the history of the study of Intelligence can be traced back to the ancient times. Humans have always been interested in the working of their own minds and developed numerous visions, theories and models to explain what I have called the human cognitive system. We seem to have a preference for formal systems to describe the human cognitive systems (or parts hereof). An example of such a formal system is logic. I have called these formal systems a computational vision. This vision sees the mind as a very complex system, following complex formulae in computing its choices. Most models that have been developed in Artificial Intelligence are based on this computational vision. These models derive their 'cognitive' abilities, not from true intelligence, but from the great computing speed and vast memory.

In the second chapter of this work, I started with the discussion of bounded rationality. This theory opposes the computational vision. Instead of seeing the mind as a perfectly logical and computational system, it takes into account the real-life boundaries of the human cognitive system. In reality, humans have to deal with limited knowledge, time and computational power. The theory of bounded rationality forms the basis of the theoretical framework that I laid down in this work.

To investigate the role of bounded rationality in communication, I looked at the generation and interpretation of nominal referring expressions. These are expressions such as *the circle, a blue circle*, etc. When humans try to communicate a certain message they have numerous ways to do so. When the utterance has been produced, there are still numerous ways to interpret the utterance. Both speaker and listener have to choose between these numerous possibilities. In a formal theory, such a choice would be modeled as a large decision tree and loads of computation would be necessary to compute the optimal choice. The key objective for my research was to test whether humans are restricted or aided by their bounded rationality in choosing the right utterance or interpretation. To cope with their bounded rationality, humans have developed fast and frugal heuristics. Fast and frugal heuristics can be seen as simple rules that describe human decision-making.

In Chapter 2.3 I discussed Optimality Theory (OT). This is a formal theory of the workings of natural language that has first been developed in the field of phonology. Nowadays several researchers have proposed OT also in the fields of syntax and semantics/pragmatics. OT describes the grammar of a language as a set of conflicting constraints that have to be resolved for each formulation and interpretation of utterances. A conflict situation is solved based on the ordering of the different constraints at stake. I have also discussed bi-directional OT. This theory incorporates the different roles a communicator has in a conversation: the roles of speaker and hearer. Bi-directional OT says that humans can use both of these roles to narrow down the possibilities for production and interpretation of utterances in conversations.

In Chapter 2.3 I have also shown a resemblance between the structure of OT and the system of fast and frugal heuristics. Both systems have a similar structure, building complex behavior on simple ranked rules (heuristic rules or constraints) and using this ranking in situations where rules are in conflict. I have shown another resemblance: bi-directional OT has a structure that resembles certain aspects of game theory. Taking the viewpoint of game theory, a conversation can be seen as an interpretation game. In this game the players interchange between the roles of speaker and hearer.

The remaining parts of the second chapter describe several theoretical concepts I needed for my research. I discussed nominal referring expressions, common ground and ACT-R. Nominal referring expressions are a clear example of natural language use in which both speaker and hearer have to determine the optimal form and meaning of expressions. The hearer needs to know which alternative forms are possible for the speaker and the speaker needs to know which alternative interpretations are possible for the hearer. The OT constraints from bi-directional OT cannot be understood without a concept of common ground. Participants of a conversation can only cooperate in communication if they understand the messages that are being exchanged and can make their own messages understandable. Finally, ACT-R is a very suitable modeling environment, in which it is possible to create a user model. ACT-R might also be a good modeling environment for further research and for the creation of actual communication agents.

My research question and hypothesis can be found in Chapter 3. The overall hypothesis of this project was that the limitations on cognitive capacities play a major role in finding the optimal strategy for choosing between different alternatives for interpretation and production of nominal referring expressions. My hypothesis was that bounded rationality actually helps finding an optimal and economic strategy. The basic concept was that participants of a conversation try to avoid the use of too many words, without taking the risk of losing the desired meaning. Also I identified several factors that influence the choice of linguistic items.

I have considered the methods for my research in chapter 4. In this chapter I laid down a set of restrictions for my experiment that can also be used for further research. I gave a description of the experimental setup: the graph completion task. In this task a participant had to solve several graphic puzzles cooperatively with an ACT-R model. A model and experiment software have been provided by Matessa. It was necessary to make several adjustments to the ACT-R model in order to make it suitable for my research for nominal referring expressions. Furthermore I described the instructions that were given to each participant and I discussed the effects of these instructions with respect to the experiment participants. At the end of chapter 4, I made several predictions for the experiment outcome based on the theoretical framework set out in chapter 2 and the hypothesis set out in chapter 3. These predictions could later be tested against the data I collected in my experiments.

The results of my experiments can be found in chapter 5. The results showed that my predictions were largely correct, notwithstanding a few remarkable effects. These effects can be taken as a lead for further research, e.g. for global strategy learning. Other suggestions can be found in chapter 5.

Other research of interest will be to investigate the role of bounded rationality in other areas of human cognition, since this project is a pilot study for the research program 'Bounded Rationality in Cognitive Strategies'. The overall research program aims at finding an integrative theory of human cognition. It is speculated that the research I performed leads the way for such an integrative theory: the view of the human cognitive system as a system of ranked and sometimes conflicting rules. We have seen evidence for such a cognitive system based on simple rules in several other areas, e.g. the Norway rats example and Petter's illusion (see chapter 2).

Complex and computational systems might be able to compute the optimal choice in complex situations, but they do not reflect the way humans seem to make their decisions. They also use more information and computational power than humans seem to do. I argue for a cognitive system based on simple rules. Such a cognitive system is not only more plausible than a complex and computational system, but it also has a great advantage over complex and computational systems: it will take less computation to make a good choice. It might not be the optimal choice, but the choice will be good enough to survive.

Finally, I wish to reflect here on the contribution that this project has made. When I started out, I expected that the emphasis would be in the actual research and experiments that I would perform. At the end of this project, sitting here behind my desk filled with papers and books, I realize that the biggest contribution that my thesis makes is not in the empirical data I collected, but in the theoretical framework that I laid down. I have combined different theoretical structures (OT, bounded rationality, simple heuristics and game theory) to create a theoretical framework that can account for the discrepancy between formal theory and actual human reasoning and thinking.

8. References

- Anderson, J.R. & Lebiere, C (1998). *The Atomic Components of Thought.* Manwah, NJ: Lawrence Erlbaum Assosiates.
- Archangeli, D & Langendoen, D.T. (1997). *Optimality Theory. An Overview.* Malden, MA: Blackwell Publishers.
- Atlas, J.D. & Levinson, S.C. (1981). It-clefts, informativeness and logical form. In P. Cole (Ed.), *Radical Pragmatics,* New York: Academic Press.
- Barlett, F.C (1932) *Remembering: a study in experimental and social psychology.* Cambridge University Press.
- Blutner, R (2000). Some aspects of optimality in natural language interpretation, *Journal of Semantics 17,* 189-216.
- Clark, H.H. (1992). *Arenas of Language Use.* Chicago: University of Chicago Press.
- Clark, H.H. (1996). *Using Language.* Cambridge, UK: Cambridge University Press.
- Cohen, P.R. (1995). *Empirical Methods for Artificial Intelligence.* Cambridge, MA: MIT Press.
- Dale, R. (1992). *Generating Referring Expressions: Construction Descriptions in a Domain of Objects and Processes.* Cambridge, MA: MIT Press.
- Dekker, P. & van Rooy, R. (2000), Bi-directional optimality theory: an application of game theory. *Journal of Semantics 17,* 217-242.
- Doyle, J (1999). Bounded Rationality. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 92-94). Cambridge, MA: MIT Press.
- Galef, B.G., Jr., (1987). Social influences of the identification of toxic foods by Norway rats. *Animal Learning & Behaviour, 15,* 327-332.
- Galef, B.G., Jr., McQuoid, L.M. and Whiskin, E.E. (1990). Further evidence that Norway rats do not socially transmit learned aversions to toxic baits. *Animal Learning & Behaviour, 18,* 199-205.
- Givon, T., (1983) Introduction. In T. Givon (Ed), *Topic continuity in discourse: Vol 3.* Amsterdam, John Benjamins.
- Grice, H. P., (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics: Vol. 3. Speech Acts* (pp. 41-58). New York: Academic Press.
- Hendriks, P & de Hoop, H. (2001). Optimality theoretic semantics, *Linguistics and Philosophy 24,* 1-32.
- Van der Hoek, W. and Verbrugge, R. (to appear in 2002). Epistemic logic: a survey. To appear in V. Mazalov and L.Petroshan (eds.), *Game Theory and Applications, Vol 8.*
- Horn, L.R. (1984). Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures. In D. Shiffrin (Ed.), *Meaning, Form and Use in Context,* (pp 11-42). Washington: Georgetown University Press.
- Horn, L (1999). Implicature. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 391-394). Cambridge, MA: MIT Press.
- Huang, Y (2000). Discourse anaphora: four theoretical models. *Journal of Pragmatics 32,* 151-176.

- Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgement under Uncertainty: Heuristics and Biases.* Cambridge, MA: Cambridge University Press.
- Kamp, H & Reyle, U (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory (Part 1).* Dordrecht, Kluwer Academic Publishers.
- Lee, B.P.H. (2001). Mutual knowledge, background knowledge and shared beliefs: their roles in establishing common ground. *Journal of Pragmatics 33,* 21-44.
- Levelt, W.J.M. (1982). Linearisation in describing spatial networks. In S. Peters & E. Saarinen (Eds.), *Processes, Beliefs and Questions* (pp. 199-220). Dordrecht: D. Reidel.
- Machina, M.J. (1987). Choice under uncertainty: problems solved and unsolved. *Journal of Economic Perspectives 1(1),* 121-154.
- Matessa, M. (2000). *Simulating Adaptive Communication.* Unpublished doctorial dissertation, Carnegie Mellon University, Pittsburg.
- Matessa, M. (2000). Interactive models of collaborative communication. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the twenty-third annual conference of the cognitive science society* (pp. 606-610). Mahwah, NJ: Erlbaum.
- McCulloch, W.S. & Pitts, W.H. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics 5,* 115-133.
- Ter Meulen, A.G.B. (2000). Optimal reflexivity in Dutch. *Journal of Semantics, 17,* 263-280.
- Partee, B.H. (1999). Semantics. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 739-742). Cambridge, MA: MIT Press.
- Prince, A. & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar* (Tech. Rep. RuCCS TR-2). Rutgers University, Center for Cognitive Science.
- Prince, A. & Smolensky, P. (1997, March). Optimality: from neural networks to universal grammar. *Science 275,* 1604-1610.
- Reinhart, T (1999a). Anaphora. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 20-22). Cambridge, MA: MIT Press.
- Reinhart, T (1999b. Binding Theory. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 86-88). Cambridge, MA: MIT Press.
- Revusky, S.H. and Bedarf, E.W. (1967). Association of illness with prior ingestion of nobel foods. *Science, 155,* 219-220.
- Roberts, C (1999). Discourse. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 231-233). Cambridge, MA: MIT Press.
- Simon, H.A. (1972). Theories of bounded rationality. In C.B. Radner & R. Radner (Eds.), *Decision and Organisation* (pp.161-176). Amsterdam: North-Holland Publishing Company.
- Simon, H.A. (1983). Alternative visions of rationality. In H.A. Simon, *Reason in Human Affairs* (pp. 7-35). Stanford, CA: University Press.
- Simon, H.A. (1997). *Models of Bounded Rationality: Vol. 3. Empirically Grounded Economic Reason.* Cambridge, MA: MIT Press.

- Smolensky, P (1999). Optimality Theory. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 620-622). Cambridge, MA: MIT Press.
- Gigerenzer, G., Todd, P. M., and the ABC Research Group (1999). *Simple Heuristics that Make Us Smart.* Oxford University Press, New York
- Todd, P. M. & Gigerenzer, G. (2000). Précis of Simple Heuristics that make us smart, *Behavioural and Brain Sciences 23,* 727-780.
- Von Neumann, J. (1958). *The Computer and the Brain.* New Haven: Yale University Press.
- Zeevat, H. (2000). The asymmetry of optimality theory and semantics, *Journal of Semantics 17,* 243-262.

Appendix 1: Instructions for the testcase

Doel van de taak:
Het gezamelijk oplossen van de incomplete figuur. Beide proefpersonen hebben een deel van de figuur bestaande uit cirkels met verbindingen ertussen. Het is zeker dat een of meer van de cirkels in beide deel-figuren gelijk zijn. Door met elkaar te communiceren over de cirkels en de verbindingen moeten de deelnemers proberen uiteindelijk beide het gehele figuur te hebben. Het experiment bestaat uit 3 fasen:
1. Vertellen / vragen
2. Figuur compleet
3. Bevestigen

Vertellen / vragen:
Voor het communiceren zijn in deze proefopstelling voor het eerste deel van het experiment enkele spelregels:
Je mag alleen iets zeggen of vragen over:
- *verbindingen tussen cirkels.* Bijvoorbeeld: “Wat zit er boven de rode cirkel?”
- *het aantal (van een bepaalde soort) cirkels.* Bijvoorbeeld: “Ik heb twee (rode) cirkels.”

Figuur compleet
Als je klaar denkt te zijn, mag je dit tiepen, of de ander vragen of hij/zij klaar is. Dit doe je pas als je denkt dat je het figuur compleet hebt. Mocht jij of de ander denken dat dit niet zo is, zeg dan dat je nog niet klaar denkt te zijn en ga terug naar **Vertellen / vragen**. Als jullie beide klaar zijn, kun je doorgaan naar **Bevestigen**.

Bevestigen
Het is de bedoeling dat jullie, op het moment dat jullie het erover eens zijn dat de figuur af is, de verschillende cirkels bevestigen. Het bevestigen moet in een nette manier gebeuren, één voor één. Dit doe je met behulp van de zogenaamde ‘matrix-notatie’. Hierin deel je je figuur in in rijen die je net als een schaakbord nummert. De locatie linksboven is dan A1, die rechts daarvan A2, die eronder B1, etc. Als volgt:

A1	A2	A3	...
B1	B2	B3	...
...	...	...	...

Bij het bevestigen gelden de volgende spelregels:
Je mag alleen iets zeggen of vragen over:
- één cirkel (immers één voor één)
- en dat in een matrix-notatie, bijvoorbeeld “Ik heb een rode cirkel op A2.”

Succes ermee!

Appendix 2: Testcase Email Conversatie

>> Start conversatie

```
Wat heb jij boven de zwarte cirkel?
- Ik heb niks boven de  zwarte cirkel.
Ik heb een kleine witte cirkel boven een grote zwarte cirkel
- Heb je ook een kleine witte cirkel links naast de grote
zwarte cirkel?
Nee, een kleine grijze links naast een kleine witte
- zit onder die kleine witte dan weer een kleine witte?
Nee, boven die kleine witte zit nog een kleine witte
- zit er iets anders onder die kleine witte dan?
Mijn grote zwarte
- Hoeveel kleine witte heb jij?
2 (twee)
- zit jouw grote zwarte ook aan de rechterkant, van de
figuur?
Bij mij zit de grote zwarte dus ook aan de rechterkant!!!
- OK, dan is die blijkbaar hetzelfde...
Heb jij iets boven je bovenste witte kleine cirkel?
- Nee, er zit wel een kleine grijze cirkel naast.
Zit er nog iets anders aan de gijze vast?
- Nee.
 Heb jij rechts van je bovenste kleine witte een grote
zwarte?
- Nee, daar zit niks.
Hoeveel kleine witte heb jij dan wel niet?
- ook 2, net als jij.
Wat zit er aan je onderste kleine witte??
- Daaraan zit aan de rechterkant de grote zwarte en verder
niets.
Zit er aan jouw grote zwarte nog iets anders dan links een
witte?
- nee
Heb jij een grote zwarte rechts van een kleine witte?
- ja dat klopt.
Ik heb onder mijn onderste kleine witte een grote zwarte
- Hoeveel grijze heb jij
twee
- Zit er een kleine grijze vast aan de onderste kleine witte?
Ja, links ervan
- Volgens mij zijn we eruit
Zit er een grijze aan de bovenste witte?
- Ja, links ervan.
Volgens mij weten we het
```

>> Einde conversatie

(De figuren bleken inderdaad overeen te komen. Experiment geslaagd!)

Appendix 3: ACT-R model

Inleiding

Dit bestand beschrijft de software die door Matessa is gemaakt voor de uitvoering van het experiment. Het doel van deze bijlage is het verduidelijken van de werking van de software, het aangeven van gemaakte wijzigingen. Daarnaast zullen er een aantal punten besproken worden die aangeven waarom er uiteindelijk gekozen is het model niet om te zetten naar het Nederlands.

Software

Bij de start van de pilot study gingen waren we van plan om het model van Matessa aan te passen voor ons eigen gebruik. Matessa heeft een soortgelijk experiment uitgevoerd zoals beschreven in (Matessa 2000a, 2000b). Dit model is gekoppeld aan een interface en de complete software bestaat uit een aantal bestanden:

`loader.lisp`	- opstart file (soort makefile)
`ACT-4.pfsl`	- act-r
`Actr-interface.lisp`	- koppelt de acties van de gebruiker aan het model
`Levelt-actr.lisp`	- de gebruikers interface
`Levelt.actr`	- het act-r model met de productieregels
`Restricted-actr.lisp`	- bepaald welke zinnen er gebruikt kunnen worden
`xx-shape-a/b.lisp`	- bestanden met de figuren voor model en proefpersoon

Het systeem start door `MCL 4.2` (Macintosh Common Lisp) op te starten en dan `loader.lisp` in te laden. Deze laadt dan vervolgens ACT-R in en de vier overige bestanden voor het experiment. Het experiment wordt gestart met het commando: `(mexp 1)`

De `xx-shape-a/b.lisp` bestanden werken als volgt: de `xx-shape-a` beschijft het figuur dat aan de proefpersoon zichtbaar is, de `xx-shape-b` beschrijft het figuur voor het model. In deze bestanden worden de figuren getekend volgens een standaard Lisp methode. Om een nieuwe figuur te maken hoef je alleen de beschrijvingen van de lijnen en cirkels aan te passen, mits je hetzelfde systeem gebruikt als Matessa. In `levelt-actr.lisp` staan verschillende functies die hiermee samenhangen.

Ik zal hier twee voorbeelden geven van het maken van een figuur, een lijn en een cirkel. NB: de schuingedrukte woorden zijn afhankelijk van defenities in de `lisp` bestanden van het model.

`(0 (7864461 7864536))` - zwarte lijn van positie 7864461 naar 7864536

`(212` *`SMALL ROUND`* `7864461)` - blauwe kleine ronde (cirkel) op positie 7864461

De kleuren:

0	*BLACK*
14485510	*RED*
212	*BLUE*
2078484	*GREEN*
16577285	*YELLOW*
16737282	*ORANGE*

De posities:

7864461	7864536	7864611
12779661	12779736	12779811

17694861	17694936	17695011

Let er bij het tekenen op dat je eerst te zwarte lijnen tekent op een positie en dan pas de cirkels, anders komen de lijnen over de cirkels te staan. De posities van de cirkels en lijnen zijn dezelfde die Matessa heeft gebruikt. Als je goed kijkt zie je dat er een systeem in zit: de eerste 4 (of 3) cijfers bepalen de verticale positie, de laatste 4 cijfers de horizontale positie. Op deze manier is het aantal locaties dus goed uit te breiden of te wijzigen.

Globale werking van het experiment

Hier zal ik kort omschrijven hoe het experiment werkt. Allereerst start je, zoals eerder vermeld, `MCL 4.2` en laadt je `loader.lisp`. De laatste zal dan de overige nodige bestanden inladen. Met het commando `(mexp 1)` start je het experiment. De `1` bepaalt met `welk xx-shape-a/b` bestanden begonnen wordt. Als het experiment klaar is (als de proefpersoon op 'done' drukt) zullen de volgende `xx-shape-a/b` bestand worden ingeladen. Met `(mexp 1)` start je dus met `01-shape-a.lisp` voor de proefpersoon.

Als dit allemaal is gedaan verschijnen er een aantal vensters op het scherm. De werking hiervan is op zich vrij duidlijk dus ik zal me hier concentreren op het venster Communication (met de voorbeeldzinnen) en het Chat window (bij opstarten nog niet in beeld). Als je een zin wil tiepen, kun je uit een van de opties kiezen in het venster Communication. Door de letter te toetsen die er voor de opties staat (dus een van de letters 'a' t/m 'u') krijg je een Chat window en een Sample window. In het Sample window zie je uit welke woorden je op dit moment kunt kiezen. Deze woorden kun je kiezen door ze in te toetsen in het Chat window. Woorden kunnen worden overgeslagen met TAB, je kunt terug met BACKSPACE en als je de verkeerde voorbeeldzin hebt gekozen kun je terug naar het venster Commnucation met ESC.

Matessa heeft in het bestand restricted-actr.lisp vastgelegd uit welke woorden je kunt kiezen per voorbeeld zin. De voorbeeld zinnen heeft hij ingedeeld naar combinaties van de inhoud van de zin en de speech act van de zin. Een `paired-state` is bijvoorbeeld een statement over een paar cirkels, zoals: "I see a red dot left of a large green dot". Zo zijn er in totaal 25 mogelijkheden voor zinnen. Echter, alle combinaties die eindigen op `-asses` zijn gelijk. Dit valt ook op in het venster Communication, waar je bij ieder onderdeel uit 4 mogelijke zinnen kunt kiezen, behalve bij 'Assesments'. De verschillende combinaties beschreven in dit bestand zijn van belang, omdat deze de basis vormen van de communicatie. De functie `choice` in dit bestand is dan ook erg belangrijk. Als je deze functie bekijkt zie je dat hier de combinatie word gevormd van de inhoud van de zin en de speech act. Deze combinaties kunnen door `levelt.actr` later gebruikt worden om te zien wat voor een soort zin er geuit is en wat voor een soort zin hierop als 'antwoord' hoor te komen. Dit gebeurt door middel van template-matching, wat terug te vinden is in de template functies in `levelt.actr`.

Als de proefpersoon eenmaal een zin heeft getiept en op ENTER heeft gedrukt, begint het model `levelt.actr` met het matchen van de productie-regels op de gangbare ACT-R wijze. Hierbij wordt eerst de zin geparsed en ingelezen, de template wordt bepaald en vervolgens kiest het model welke actie gewenst is. Als alles goed gaat kiest het model de goede actie (bijvoorbeeld vertellen hoeveel groene cirkels hij heeft als er naar het aantal groene cirkels gevraagd wordt) en vormt een zin. Het model is vrij snel dus er is een aftel-mechanisme ingebouwd, dat aftelt tot er een 'menselijke' tijd met het antwoord is gewacht. Als er afgeteld is, zal het model het antwoord in het Chat window laten verschijnen en is de proefpersoon weer aan de beurt.

Problemen en aanbevelingen

In eerste instantie hadden we verwacht dat het vertalen veel tijd zou kosten, maar niet al te moeilijk zou zijn. De woorden moesten van het Engels naar het Nederlands en een aantal keren (zoals bij vraagzinnen) zou de volgorde anders moeten. Er zijn echter een aantal problemen ontstaan waar we geen rekening mee hebben gehouden.

1. Zoals hierboven vermeld heeft Matessa 25 mogelijke zinsconstructies bedacht voor het communiceren. Hierop heeft hij zijn experiment en zijn model gebaseerd. Deze zinnen komen echter niet helemaal overeen met de gangbare Nederlandse zinnen. Een zin als "Let's see how many green dots we have" is in het Nederlands niet gebruikelijk. Een proefpersoon zou hier gebruik maken van "Hoeveel groene cirkels heb jij", een van de andere opties. De door Matessa gemaakte opties zijn een essentieel basisonderdeel van het programma, omdat hier alle templates en productieregels op zijn gebaseerd alsmede de mogelijke zinnen en woorden.
2. Bij het vertalen kun je niet altijd eenduidig te werk gaan. De meeste woorden kun je direct vertalen, maar sommige woorden in het Engels moet je vertalen met meerdere woorden in het Nederlands en andersom. "I'm confirming" wordt bijvoorbeeld "Ik bevestig", maar "I'm done" wordt "Ik ben klaar". In de productie-regels is er echter maar 1 regel voor "I'm" en in het Nederlands zijn hier dus meerdere regels voor nodig.
3. Een aantal constructies zijn door Matessa hard ingebakken. Hiermee bedoel ik dat in sommige productie-regels letterlijk woorden zijn opgenomen als voorgangers van woorden (previous-word) of als huidige argumenten (arg1, arg2 of relation). Dit hoeft niet altijd een probleem te zijn, maar als de woordvolgorde veranderd, of als de zinsconstrucite in het Nederlands zelfs compleet anders is, kunnen deze produtieregels niet blijven staan. Er moeten dan nieuwe productieregels komen.

Kortom: de code zoals Matessa die heeft gemaakt is volgens mij niet geschikt om te vertalen naar het Nederlands. De voornaamste reden is dat er in het Nederlands gewoonweg andere zinnen worden gebruikt (waarschijnlijk minder) dan in het Engels. Hierbij komen de problemen met woordvolgorde, zinsconstructies en dergelijke bij het vertalen.

Het is dus waarschijnlijk noodzakelijk om een nieuw model te gaan maken. Bij het maken van dit model moet aandacht besteed worden aan de volgende zaken:

Allereerst zal er een onderzoek gedaan moeten worden naar de Nederlandse zinnen zoals die voorkomen in vergelijkbare situaties. Dit kan door een studie te doen met een aantal proefpersonen die dezelfde taak krijgen (het figuur compleet maken) maar dan via een chat systeem zonder de huidige restricties. Het is dan wel belangrijk om restricties op te leggen aan de soorten uitspraken die er gedaan kunnen worden. Proefpersonen zouden het dan moeten houden bij vragen en stellingen over één cirkel of twee cirkels die met elkaar verbonden zijn. Bovendien zou in één uiting of een antwoord of een vraag gesteld moeten kunnen worden. (Zie hiervoor ook de bijlage **Figuur Taak**, die ik geschreven heb voor mijn eigen test in het Nederlands. Hierbij moet opgemerkt worden dat het waarschijnlijk verstandig is geen voorbeeldzinnen te noemen om de proefpersonen niet te beïnvloeden).

Als dit onderzoek gedaan is, kan er een aantal 'standaardzinnen' bepaald worden zoals die gebruikt worden door proefpersonen als ze vrij kunnen tiepen. Deze standaardzinnen kunnen vervolgens vertaalt worden naar het model en op basis hiervan kan er dan een nieuw model gemaakt worden. Bepaalde onderdelen van het oude model van Matessa kunnen dan zeker nog gebruikt worden, zoals de interface en het gebruik van templates.

Referenties

- **(Matessa, 2000a)**
  Matessa, M. (2000). *Simulating Adaptive Communication.* Unpublished doctorial dissertation, Carnegie Mellon University, Pittsburg.
- **(Matessa, 2000b)**
  Matessa, M. (2000). Interactive models of collaborative communication. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the twenty-third annual conference of the cognitive science society* (pp. 606-610). Mahwah, NJ: Erlbaum.

Appendix 4: Other translation problems

Normal statements such as

A large green circle is above our small yellow circle

can be translated easily into Dutch:

Een grote groene cirkel is boven onze kleine gele cirkel.

This was done successfully, but other types of sentences posed more serious problems. Question sentences for example have a different construction in English as in Dutch.

What is above our large green circle?
Wat is er boven onze grote groene cirkel?

The addition of the word *er* in Dutch poses serious problems, because the ACT-R production rules were written specifically for the English construction. Specific ACT-R production rules could parse words of the *is* type (such as *is*, *appears*, etc), but weren't capable of parsing the construction including the Dutch *er*. Another example is the English use *I'm*. In Dutch there are two different constructions depending on the use of *I'm*:

1.a *I'm confirming ...*
1.b *Ik bevestig ...*

2.a *I'm done ...*
2.b *Ik ben klaar ...*

In the first sentence the translation into Dutch changes the infinitive *confirming* into the first person *bevestig*. In the second sentence the translation into Dutch does not make this transition from infinitive into first person verb. Here the structure of the sentence stays intact. In fact, this is a problem that results from the response mechanism with the recombination sentence types. Had the model been based on a grammar, it would probably have been possible to fix this problem.

Another problem was the difference between the ZERO expressions in English and Dutch. In Dutch using a ZERO expression has a economical advantage over the use of a ZERO expression in English. Referring to a blue circle in English using a ZERO expression would be:

The blue one

In Dutch this would be:

De blauwe

Here it is clear that in Dutch the use of a ZERO expression is more economical.

Appendix 5: ACT-R regels

Nieuwe regels die zijn toegevoegd om het 'ontkoppelen' van de dimensies van de cirkels mogelijk te maken. De regels zijn alleen bij naam genoemd.

```
4 dimensies in totaal
========================
ground-middle-color-size-shape
ground-middle-color-size-shape2
ground-topmost-color-size-shape
ground-topmost-color-size-shape2
ground-bottommost-color-size-shape
ground-bottommost-color-size-shape2
ground-leftmost-color-size-shape
ground-leftmost-color-size-shape2
ground-rightmost-color-size-shape
ground-rightmost-color-size-shape2

3 dimensies in totaal
=============================
ground-middle-color-size
ground-middle-color-size2
ground-topmost-color-size
ground-topmost-color-size2
ground-bottommost-color-size
ground-bottommost-color-size2
ground-leftmost-color-size
ground-leftmost-color-size2
ground-rightmost-color-size
ground-rightmost-color-size2

ground-middle-color-shape
ground-middle-color-shape2
ground-topmost-color-shape
ground-topmost-color-shape2
ground-bottommost-color-shape
ground-bottommost-color-shape2
ground-leftmost-color-shape
ground-leftmost-color-shape2
ground-rightmost-color-shape
ground-rightmost-color-shape2

ground-middle-shape-size
ground-middle-shape-size2
ground-topmost-shape-size
ground-topmost-shape-size2
ground-bottommost-shape-size
ground-bottommost-shape-size2
ground-leftmost-shape-size
ground-leftmost-shape-size2
ground-rightmost-shape-size
ground-rightmost-shape-size2

Bijgemaakte regels om het grounden van cirkels zonder
locatiegegevens mogelijk te maken
============================
ground-color-shape-size2   (hoogste expected gain)
```

```
ground-color-shape2        (expected gain met :r 0.7)
ground-color-size2         (expected gain met :r 0.7)
ground-shape-size2         (expected gain met :r 0.7)

4 dimensies, andere soort regels
===============================
get-topmost-color-shape-size
get-bottommost-color-shape-size
get-leftmost-color-shape-size
get-rightmost-color-shape-size
get-middle-color-shape-size

3 dimensies in totaal
==========================
get-topmost-color-size
get-bottommost-color-size
get-leftmost-color-size
get-rightmost-color-size
get-middle-color-size

get-topmost-color-shape
get-bottommost-color-shape
get-leftmost-color-shape
get-rightmost-color-shape
get-middle-color-shape

get-topmost-size-shape
get-bottommost-size-shape
get-leftmost-size-shape
get-rightmost-size-shape
get-middle-size-shape


get-regels voor communicatie
===========================
get-color-shape-size-topic
get-color-shape-size-fill
get-color-shape-size-match
get-color-shape-size-nomatch-topic
get-color-shape-size-nomatch
get-color-shape-size-nomatch-conf
get-color-shape-size-any
get-color-shape-size-failure
get-color-shape-size-failure-conf-topic
get-color-shape-size-failure-conf

get-color-shape-topic
get-color-shape-fill
get-color-shape-match
get-color-shape-nomatch-topic
get-color-shape-nomatch
get-color-shape-nomatch-conf
get-color-shape-any
get-color-shape-failure
get-color-shape-failure-conf-topic
get-color-shape-failure-conf

get-color-size-topic
```

```
get-color-size-fill
get-color-size-match
get-color-size-nomatch-topic
get-color-size-nomatch
get-color-size-nomatch-conf
get-color-size-any
get-color-size-failure
get-color-size-failure-conf-topic
get-color-size-failure-conf

get-shape-size-topic
get-shape-size-fill
get-shape-size-match
get-shape-size-nomatch-topic
get-shape-size-nomatch
get-shape-size-nomatch-conf
get-shape-size-any
get-shape-size-failure
get-shape-size-failure-conf-topic
get-shape-size-failure-conf

==============================================================
=====
==============================================================
=====
Hieronder staan oude regels die al aanwezig waren in het
model, maar waar aanpassing nodig was om het geheel goed te
laten werken, o.a. voor het 'ontkoppelen' van de dimensies.
Deze zijn aangepast door extra toevoegingen, zoals 'shape
nil' en 'color nil' als er alleen stond 'size =size'
Anders zouden deze regels ook gaan vuren als size is gevuld,
maar een ander slot
ook en je wil dat ze alleen gaan vuren als alleen de size
bekend is.

Locatie regels
==========================
ground-above
ground-above2
ground-below
ground-below2
ground-left
ground-left2
ground-right
ground-right2

2 dimensies
==========================
ground-middle-color
ground-middle-color2
ground-topmost-color
ground-topmost-color2
ground-bottommost-color
ground-bottommost-color2
ground-leftmost-color
ground-leftmost-color2
ground-rightmost-color
ground-rightmost-color2
```

```
ground-middle-shape
ground-middle-shape2
ground-topmost-shape
ground-topmost-shape2
ground-bottommost-shape
ground-bottommost-shape2
ground-leftmost-shape
ground-leftmost-shape2
ground-rightmost-shape
ground-rightmost-shape2

ground-middle-size
ground-middle-size2
ground-topmost-size
ground-topmost-size2
ground-bottommost-size
ground-bottommost-size2
ground-leftmost-size
ground-leftmost-size2
ground-rightmost-size
ground-rightmost-size2

1 dimensie zonder locatie
=========================
ground-color2
ground-size2
ground-shape2

2 dimensies
====================================================
get-topmost-color
get-bottommost-color
get-leftmost-color
get-rightmost-color
get-middle-color

get-topmost-shape
get-bottommost-shape
get-leftmost-shape
get-rightmost-shape
get-middle-shape

get-topmost-size
get-bottommost-size
get-leftmost-size
get-rightmost-size
get-middle-size

1 dimensie
===========================================
get-color-topic
get-color-fill
get-color-match
get-color-nomatch-topic
get-color-nomatch
get-color-nomatch-conf
get-color-any
```

```
get-color-failure
get-color-failure-conf-topic
get-color-failure-conf

get-size-topic
get-size-fill
get-size-match
get-size-nomatch-topic
get-size-nomatch
get-size-nomatch-conf
get-size-any
get-size-failure
get-size-failure-conf-topic
get-size-failure-conf

get-shape-topic
get-shape-fill
get-shape-match
get-shape-nomatch-topic
get-shape-nomatch
get-shape-nomatch-conf
get-shape-any
get-shape-failure
get-shape-failure-conf-topic
get-shape-failure-conf
```

Appendix 6: Uitleg Experiment

Hieronder staat een en ander uitgelegd over het experiment. Lees deze uitleg goed door voor het experiment begint! We zullen ook nog 1 testronde met je doen, zodat je kunt wennen aan de manier van werken in dit experiment.

Doel
In het experiment krijg je een deel van een figuur te zien op je scherm, bestaande uit cirkels met verbindende lijnen. Het doel van het experiment is, om samen met je partner de gehele figuur samen te stellen. Je hebt allebei namelijk een deel van de figuur, waarbij één of meer van de cirkels overlappen. Er is één cirkel waarvan je van tevoren weet dat jij en je partner die gemeenschappelijk hebben. *Dit is de cirkel met de zwarte stip erin.* Aan het einde van elke puzzel moet je stuk voor stuk de cirkels bevestigen. Hier kun je punten mee verdienen.

Regels
- Om onduidelijkheid te voorkomen mag je voor de gemeenschappelijke cirkel het woordje 'our' (= onze) gebruiken. Dit woord mag je dus niet voor andere cirkels gebruiken.
- Relaties als 'above' en 'left of' gelden alleen als er een verbinding is tussen twee cirkels. 'Above' geldt dus niet als een cirkel schuin boven een andere cirkel staat.
- Het belangrijkste doel van het spel is dat de figuur uiteindelijk compleet wordt. Daarnaast is het de bedoeling dat je dit zo snel mogelijk doet. Kortom: het doel is een zo hoog mogelijke score in een zo kort mogelijke tijd.
- Je moet altijd de waarheid spreken: als je partner een vraag stelt, moet je deze naar waarheid beantwoorden.
- Je moet bovendien altijd doen wat de ander zegt: een speler moet bijvoorbeeld altijd een object plaatsen op een bepaalde plek als de ander dat zegt.

De vensters op het scherm
Drawing Pad: het veld met jouw deel van het figuur erin linksboven in beeld.
Communication: de lijst met voorbeeldzinnen van A tot U rechts van het Drawing pad.
Score: het kleine venstertje boven het Communication venster.
Chat window: het kleine venster waar je zinnen in kunt typen. Dit is nog niet in beeld
Zinnen-bouwer: het venster waar de mogelijke woorden die je kunt typen in verschijnen. Dit is ook nog niet in beeld.

Communication
Toets een letter ['a' t/m 'u'] om een type zin te selecteren.
Als je een verkeerde voorbeeldzin hebt gekozen kun je terug naar het menu met ESCAPE.
Je kunt in het Chat window een van de woorden typen die in de Zinnen-bouwer verschijnen. Je kunt dan alleen die woorden typen. Zodra je het woord hebt getiept, komt er een spatie achter het woord en kun je een tweede woord typen.
Als er iets fout gaat met typen, kun je gewoon BACKSPACE gebruiken.

Je kunt altijd woorden overslaan door op de TAB-toets te drukken
Als je je zin wilt versturen, druk je op ENTER. Dit kan even duren.

Drawing pad
Deze heeft 4 functies:
Cirkels toevoegen aan je figuur (add)
Cirkels verwijderen uit het figuur (erase)
Cirkels bevestigen samen met je partner (confirm)
En als het bevestigen klaar is, kun je op 'Done' drukken (onderaan). Dit is onherroepelijk, dus als je hierop drukt is het experiment afgelopen met het huidige figuur. Dus pas op 'Done' drukken als je klaar bent met bevestigen!!

Het werkt als volgt: als je een object wilt toevoegen aan je figuur selecteer je eerst de juiste eigenschappen bovenaan in het venster. Dus bijvoorbeeld: <red> <small> <circle> en je vinkt de knop 'add' aan.

Dan klik je op een van de bestaande objecten. Hierin verschijnt dan een wit stipje. (Je kunt nu ook nog een ander object aanklikken en dan komt het witte stipje daar te staan.) Als je nu op een plek klikt naast, boven of onder het object met de witte stip, komt daar je nieuwe object. (In dit geval dus een kleine rode cirkel.)

Als je een object hebt geplaats dat je weer wilt weghalen, dan kun je dit doen door het hokje 'erase' aan te vinken (in plaats van het hokje 'add' dat standaard aangevinkt is). Als je dan op een cirkel klikt, zal deze verdwijnen uit je figuur.

Aan het eind van het experiment moet je samen met je partner alle objecten langslopen en bevestigen. Dit doe je door het hokje 'bevestigen' aan te vinken en dan te klikken op het object dat je wilt bevestigen, of het object waarvan je denkt dat je partner het wil bevestigen. Je kunt elk object maar één keer bevestigen. Als dit goed gaat krijg je er 10 punten bij, als dit fout gaat, gaan er 10 punten af. Dit kun je zien in het Score venster. Het score venster geeft het bericht 'waiting for you...' als jij moet bevestigen en het bericht 'waiting for partner...' als je partner moet bevestigen.

Pas als je helemaal klaar bent met bevestigen, kun je op de knop 'Done' drukken onderin het venster Drawing pad. Er zal dan een nieuwe figuur komen in de Drawing Pad en het experiment begint weer van voren af aan.

Dat was de uitleg.
Succes met de puzzels!