Bi-Directional Optimality Theory: An Application of Game Theory

Paul Dekker and Robert van Rooy

Abstract
Optimality Theory catches on in linguistics, first in phonology, then in syntax, and recently also at the semantics / pragmatics interface. In this paper we point at some parallels between principles employed in optimality theoretic interpretation, and notions from the well-established field of Game Theory. Optimality theoretic interpretation can be defined as what we call an "interpretation game", and optimality itself can be viewed as a solution concept for a game. More in particular, optimality can be characterized in terms of the game-theoretical notion of a "Nash Equilibrium".

1 Introduction

If John says that OTS is *possibly* right, we can infer from this that he thinks it is *not* obviously, or *necessarily* right. What kind of inference is this? Suppose that from *Possibly A* we can infer *semantically* that it is possible that A is false. By this assumption we can easily account for the above inference, but we can no longer account for the fact that we might appropriately say *OTS is possibly right, if not necessarily.* The latter example makes clear that the above inference to the possibility that OTS is wrong cannot be conventionally associated with all sentences in which the sentential clause *OTS is possibly right* occurs. But how then should we account for the intuition that we can conclude that OTS might be wrong from what John says? Following Grice (1975), it has become a common practice in the area of pragmatics to distinguish what is *said* by the speaker's use of a sentence (the *semantic* or *truth-conditional* meaning of a sentence), and what is *meant* by it on a particular occasion. Thus conceived, pragmatics is concerned with the study of what is meant by an utterance above its semantic, or truth-conditional, content by taking into account the issue whether the utterance is *appropriate* in its conversational context, i.e. with respect to the (common) beliefs and intentions of the participants of the conversation. The main motivation for this division of labour between semantics and pragmatics is to keep the semantics as simple as possible; it allows us to determine the semantic content of a sentence in a *compositional* way based on its syntactic structure, without making reference to the attitudes of speakers and hearers.

Following Gazdar (1979), the following general pipe-line architecture of the semantics / pragmatics interface has emerged

1. what is said by a (declarative) sentence, its semantic content, is equated with its truth-conditions
2. truth-conditional content can be determined in a rather simple way compositionally without making reference to either what is (or could be) pragmatically implicated by what is said, or the attitudes of the participants of the conversation
3. to determine what is pragmatically implicated we can, and have to, make use of the truth-conditional content of the sentence; what is potentially implicated might be *overruled*, or *canceled*, if it conflicts with what is semantically entailed, as in our above example *OTS is possibly right, if not necessarily.*

Thus, according to Gazdar, the semantics / pragmatics interaction goes only one way; although what is pragmatically presupposed or implicated might depend on the semantic content of the sentence, semantics is *autonomous* from pragmatics.

It seems clear to us that this strong Gazdarian picture of the interface must be wrong for the following reason: not only what is pragmatically implicated depends on the attitudes of the participants of the conversation, but this might also the case for the truth-conditions that a sentence has. Pragmatic notions like *appropriateness*, *expectation/naturalness* and *relevance* are both used to determine what is conversationally implicated *and* to determine what is asserted by a sentence. It is clear that this dependence of what is said, or asserted, on pragmatic notions undermines the goal to determine the truth-conditions of sentences in a compositional way. For an illustrative example, let us consider briefly the process of anaphora resolution for a sentence like <u>*He*</u> *is tall.* It is clear that this sentence is highly underspecified or ambiguous; in different contexts the pronoun might refer to different individuals. Its resolution implies reference to such things as focus (Sidner, 1983), the syntactic position (subject/non-subject) of the antecedent (Grosz et al, 1995), but also to the scenarios/prototypical situations involved (e.g. Sanford & Garrod, 1981). Although the meaning of the sentence is highly context-dependent, the sentence has a more constant meaning too; we might say that in all contexts the pronoun refers to the most salient male individual in that context. A Gazdarian might then propose to represent this contextual information in a more or less objective way, without referring to the attitudes of the agents. What is the most salient individual in a context? For some contexts we can give rather objective criteria. For instance, it seems clear that when we utter the above sentence in the context where *Bill is next to John* has just been uttered, the pronoun will refer to Bill, but when the foregoing sentence would have been *John is next to Bill*, the pronoun would refer to John. The objective criterium in this case is that the (individual denoted by the) subject of a preceding sentence is more salient than the (individual denoted by the) object. But now consider the following discourse: *Bill tickled John. <u>He</u> squirmed.* According to the above rule the pronoun should refer to *Bill*. It is clear, however, that according to its most reasonable interpretation the pronoun does not refer to Bill, but to John. Why? Because we assume that it is the tickled person who has reason to squirm; the assertion that John squirmed is more in accordance with the expected scenario triggered by the previous sentence than the assertion that it is Bill who squirmed. We conclude that the speaker asserted that John squirmed, i.e., the constraint that the pronoun refers to the most salient person in its context of interpretation is *overruled* by the constraint that demands that what is said should be natural in its context of interpretation, i.e., in accordance with the relevant scenario. The triggered scenarios depend on world-knowledge and expectations of the participants of a conversation, which suggests that the relevant contextual parameters cannot be given without making reference to the attitudes of the speakers. But now we are running ahead of ourselves. For we might think of representing the relevant contextual parameter in the context of interpretation of the sentence in which the pronoun occurs as an 'objective' salient order, when we allow with Lewis (1979) for a rule of *accommodation* of comparative salience. In principle this is feasible, but note that in this case it is the process of accommodation that is governed by notions like *appropriateness*, *naturalness* or *relevance* that cannot be described without making reference to the attitudes of agents. Notice that according to this variant the relevant contextual parameter that helps to determine what is said (its truth conditions) by an utterance, the salience ordering, crucially depends on the utterance itself; whether and how the salience order should be accommodated depends on what would have been said by this utterance according to the different possible salience orderings. Observe also that in this variant some constraints can be overruled by our general pragmatic notions; in this case not that a pronoun should refer to the most salient individual in its context, but rather that the salience order determined after the interpretation of the first sentence of a discourse will function as the relevant salience order to interpret the anaphoric pronouns of the following sentence.

The above example shows that we cannot systematically determine the semantic content of a sentence in a *compositional* way based on its syntactic structure, without making reference to the attitudes of speakers and hearers, if we equate the semantic content of a sentence with its truth-conditions. So what should we do? Give up compositionality, or give up the assumption that what should be determined compositionally are the truth-conditions of a sentence? The former, radical, option would result almost surely in giving up the distinction between semantics and pragmatics, as has been proposed in the old days of generative semantics. According to the latter option, compositional semantics still has a role to play. However, the semantic content of a sentence is not fully determined and doesn't give rise to clearcut truth-conditions, it is left *underspecified.*

We have only discussed pronouns above, but similar remarks can be, and have been, made for the interpretation of other context-dependent constructions like modals (Kratzer, 1979), presuppositions (van der Sandt, 1992), quantifier scope (Parikh, 1993), tenses (Asher & Lascarides, 1994), adjectives (Blutner, 1998), and quantified constructions (Hendriks & de Hoop, to appear). For all those cases it has been proposed that what should be determined compositionally should be left rather underspecified, and that to determine the actual truth-conditions of a sentence we have to rely on constraints motivated by principles of rational communication as given, for instance, by Grice's maxims of conversation. This results, obviously, in a new formulation of the semantics / pragmatics interface.

2 Optimality Theoretic Interpretation

Recently, various phenomena on the semantics / pragmatics interface, like the ones discussed above, have been given an optimality theoretic formulation (Blutner, Hendriks and de Hoop, de Hoop and de Swart, Jäger, Zeevat). In this section, and in section 4, we give a short overview of the various types of analyses that have been proposed, and illustrate these by means of a few examples.

2.1 One-Dimensional Optimality

According to the proposed application of Optimality Theoretic principles by de Hoop & de Swart (to appear) and Hendriks & de Hoop (1999) to the theory of interpretation, what compositional semantics gives us is a radically underspecified notion of meaning represented by a possibly infinite set of interpretations of a well-formed syntactic structure. In addition, optimality theory gives us a ranked set of constraints which allow us to select the optimal interpretation associated with a particular syntactic structure. These constraints should of course be as general as possible, and also the rankings between those constraints should, if possible, be valid for a wide range of languages, based on general principles of rational communication.

In order to illustrate how things might work out in such a theory, consider again the example which we discussed above with an anaphoric pronoun. The example is of the form *aRb. <u>He</u> is P*, where in the first sentence a and b are both names for male individuals. Discourses of this form are potentially ambiguous, or underspecified, because the pronoun might refer back to either a or b. But we can say something more; on the basis of empirical data we might observe that the pronoun will typically refer back to the subject expression, i.e. a. We can state this observation explicitly in a constraint. This constraint is very particular, but we might embed this particular constraint within a more general one, if we make use of the notion of comparative salience. In whatever way we do this, the important point is that the relevant constraint should *not* be too *hard*; in some circumstances it might be overruled. In the above discussed discourse *Bill tickled John. <u>He</u> squirmed*, for instance, it doesn't seem *natural* to state that Bill squirmed after the first sentence. Because it seems reasonable, with an eye upon the communicative aims, to assume that the constraint on naturalness is more important than the constraint on salience, the constraint that in our case demands that the pronoun should refer to the subject expression of the previous sentence becomes overruled. Thus, although pronouns are meant to refer back to subject expressions of previous sentences, this will only result in an *optimal interpretation* in case the stronger constraint of naturalness is also met.

Another example, discussed in Hendriks and de Hoop, is the following:

(1) Often when I talk to a doctor$_i$, the doctor$_{\{i,j\}}$ disagrees with him$_{\{i,j\}}$.

In the interpretation of this example two constraints are at work:

(B) If two arguments of the same semantic relation are not marked as being identical, interpret them as being distinct

(DOAP) Don't Overlook Anaphoric Possibilities

In example (1), the two constraints have conflicting effects. If (DOAP) is fully satisfied, that is, if both "the doctor" and "him" are interpreted as anaphoric upon the "a doctor", then (B) is violated. And if (B) is satisfied, then at least either "the doctor" or "him" remains unresolved. Intuitively, this seems the best solution, and Hendriks and de Hoop therefore use this example to show that constraint (B) is harder than (DOAP). The (DOAP)-principle can be overruled in order to satisfy (B), and the 'optimal' interpretation is that either "the doctor" and not "him" is anaphoric upon the antecedent "a doctor", or the pronoun and not the definite description is.

So far we have sketched an optimality theoretic formulation of only one of the two types of pragmatic inferences which we discussed in the first section of this paper. So how should we account for the case with which we began our story: the *scalar implicature* from $\Diamond A$ to $\neg\Box A$? Our intuitive explanation for this implicature was that the speaker did not think it is necessary that OTS is right, because otherwise he would have said so, i.e., he would have used *another expression*. It is not entirely clear how to account for this reasoning in terms of the above sketched one-dimensional search for optimality where the input is given by single syntactic structure, and no reference is made to alternative expressions that the speaker might have used. Blutner (ms) has recently argued that an account of scalar implicatures requires us to take into consideration what the speaker *could* have said, and proposed to go from a one-dimensional, to a two-dimensional search for optimality.[1] This two-dimensional view was mainly motivated by a reduction of Grice's maxims of conversation to two principles.

2.2 The Q- and I-Principles

In his seminal paper on Logic and Conversation, Grice (1975) tried to account for so-called pragmatic inferences by making use of four maxims of conversation: the maxims of *quality*, *quantity*, *relation*, and *manner*. More

1. The idea to compare not only different outputs with each other to determine the optimal interpretation, but also to take different inputs into account, can be traced back to Prince and Smolensky's (to appear) principle of Lexicon Optimization (section 9.3). A bi-directional view on optimality plays implicitly also an important role in the OT learning algorithm (Tesar & Smolenksy, to appear). According to this algorithm each piece of positive evidence (structural description) about the correct ordering of constraints brings with it a body of implicit negative evidence; the chosen description is preferred to the given competitors.

recently, some attempts have been made to reduce and explicate these maxims to some more principled rules of, or constraints on, rational behavior in communication. Valuable contributions in this direction have been made especially by Atlas & Levinson (1981) and Horn (1984), who seek to reduce the maxims of quantity, relation and manner to the following two principles; the Q-principle, (implementing Grice's first maxim of quantity), which advises the speaker to say as much as he can to fulfill his communicative goals, and the I-principle (called R-principle by Horn (1984), and implementing the rest of the Gricean maxims except for quality), which advises the speaker to say no more than he must to fulfill his communicative goals. By means of the I-principle we can explain, for instance, why in many contexts we can use (short, and thus efficient) pronouns to refer to individuals, instead of long eternal definite descriptions, and it can also help to explain why in many cases the conjunctive connective *and* gives rise to a temporal, or even causal, interpretation. The Q-principle is responsible for the so-called *scalar implicatures*, and makes essential reference to *alternative expressions* the speaker could have used.

Although both principles have the effect that the hearers can conclude more from the utterance than what is explicitly said by it, the strenghthenings due to the I and Q principles typically go in *opposite directions.* As a result, the two principles sometimes advise the speaker to do opposite things, and thus we would expect that the hearers sometimes don't know what to make of the utterance. For instance if you say *John was able to solve the problem*, I can conclude by means of the I-principle that John actually solved the problem, while the Q-principle gives rise to the opposite conclusion that John actually did *not* solve the problem. (For otherwise you should have said he did so.) Horn (1984), following Zipf (1949), gives an interesting motivation for why the I- and Q-principles seem to give rise to opposite conclusions. He argues that the principles can be seen as representations of rational goals of competing forces to *minimize their efforts*: The I-principle represents the *speaker's* goal to minimize the effort to *communicate* as much as possible, while the Q-principle can be seen to represent the *hearer's* goal to minimize his effort to *understand*.[2]

Looking at both principles from a minimization point of view has the effect that the I-principle and the Q-principle should be seen from two

2. Notice the resemblance with Sperber & Wilson's (1986) Relevance Theory, according to which meaning – optimal relevance – can be thought of as a balance between the two competing forces of *maximization* of contextual effect and *minimizing* of processing effort.

different perspectives: the I-principle from the speaker's perspective, and the Q-principle from the hearer's perspective. Interestingly, the principles can be viewed, equivalently it seems, from a *maximization* point of view when we switch roles. That is, an I-maxim requiring a cooperative speaker to say no more than needed, will make a rational hearer to get as much as possible out of which the speaker says, that is, in such a cooperative setting, the I-principle relates to a *hearer's* goal to maximize the *relevance*, or informativity, of a given utterance. Conversely, the Q-principle, advices the speaker to maximize his contribution to the goal of being as *informative* as he can. (As it indeed was upon Grice's formulation.) The two points of view thus collaborate to achieve two mutually dependent goals of the interlocutors: to maximize the cooperative and mutual goal of informativity, and to minimize individual efforts.

2.3 Two-Dimensional Optimality Theoretic Interpretation

Blutner (1998, 1999) has recently given the I- and Q-principle a slightly different formulation such that the Gricean maxims can be seen as being part of a *two*-dimensional optimality theoretic framework of disambiguation. The I-principle is formulated much like it was above from a maximization point of view, and helps to select the most coherent, or relevant, interpretation. This principle corresponds to the one-direction view on optimality theoretic interpretation as proposed by Hendriks & de Hoop (to appear) and De Hoop & De Swart (to appear), which, exclusively, adopt the *hearer's perspective* on disambiguation. What is interesting is that Blutner also implements the Q-principle within an Optimality Theoretical framework, thereby also taking the *speaker's perspective* into account. Where the I-principle compares different possible interpretations for the same syntactic expression, the Q-principle compares different possible syntactic expressions that the speaker could have used to communicate the same meaning. The interesting feature of Blutner's formulation of the Q-principle within two-dimensional OT is that although it compares alternative syntactic *inputs* to one another, it still helps to select the optimal meaning among the various possible *outputs* of the single actual syntactic input given, by acting as a *blocking mechanism*.[3] The strong version of Blutner's two-dimensional OT can be formulated as

3. Boersma (1998) has recently made a similar move in phonology. He argues that sound structures reflect an interaction between the articulatory and perceptual principles of efficient and effective communication; the speaker-oriented principle of *minimization of articulatory effort* and the hearer-oriented principle of *minimization of perceptual confusion*.

follows (we here relate pairs (r, m) of possible representations (r) and meanings (m), by means of an ordering relation '$>$', 'being more efficient'):

(2) Two-dimensional OT (Strong Version) a representation-meaning pair (r, m) is *optimal* iff it satisfies both the Q- and the I-principle, where:
  (Q) (r, m) satisfies the Q-principle iff there is no other pair (r', m) such that $(r', m) > (r, m)$
  (I) (r, m) satisfies the I-principle iff there is no other pair (r, m') such that $(r, m') > (r, m)$

How does this blocking due to the Q-principle work? Consider the scalar implicature again from *Possibly A* to *Not necessarily A*. Let us suppose that the speaker knows all about the possibility of A, and that he has the opportunity to say *Possibly A* ($\Diamond p$), *Necessarily A* ($\Box p$) and the negation of these modal possibilities ($\neg\Diamond p \equiv \Box\neg p$ and $\neg\Box p \equiv \Diamond\neg p$, respectively). Let us also assume, as seems quite natural, that $\Box p \models \Diamond p$. Given these assumptions, the speaker knows that only one of three logical possibilities obtains:

(i) that $\Box p$ (and, hence, $\Diamond p$),
(ii) that $\Diamond p \wedge \Diamond\neg p$ (so $\neg\Box\neg p \wedge \neg\Box p$), or
(iii) that $\Box\neg p$ (i.e., $\neg\Diamond p$ and $\Diamond\neg p$).

Furthermore, there is a common preference to communicate as much as possible, that is, a preference for (i) and (iii) over (ii). In this situation, saying *Possibly A* ($\Diamond p$) implicates $\Diamond\neg p$. For if the speaker had information to the effect that $\neg\Diamond\neg p \equiv \Box p$, he would have said *Necessarily A*, which is more informative. As he has not done so, and as long as there is no reason to suppose otherwise, the hearer is entitled to infer $\Diamond\neg p$. So, although the sentence *Possibly A* is logically consistent with both *Necessarily A* and *Possibly not A*, the first is *blocked* (by the Q-principle), because of the existence of an alternative syntactic form that would express that meaning in a more efficient way.

3 Game Theory and Strong Optimality

The ranking and judging of representations and meanings in optimality theoretic interpretation has a structure which resembles principles developed in the well-investigated field of Game Theory. In this section we present a game-theoretical formulation of Blutner's notion of optimality. (For an in-depth introduction to game theory, cf., e.g., Osborne and Rubinstein 1994). The first section presents an introduction to some of the basics of Game

Theory, in particular to that of a strategic game. In the next subsection we present the notion of a 'Nash Equilibrium', a renown solution concept in Game Theory. In the third subsection we then show how optimality theoretic interpretation can be given a formulation in terms of an interpretation game, and that Blutner's concept of optimality corresponds to precisely this concept of a Nash Equilibrium.

3.1 A Formal Definition of Games

In Game Theory, a 'strategic game' is the formal rendering of a game which can be played with a specific number of players, who can play various roles in the game. In strategic games it is assumed that the players all make one choice at the beginning of the game. The players (simultaneously) choose a strategy, and then they play the game, each according to the strategy chosen. It is assumed that the players know what options are available to them and to the other players, and what are the outcomes of the game if they know the actions chosen.

A strategic game is formalized as a triple $\langle N, (A_i), (\geq_i)\rangle$ which consists of a set of players N, and, for each player $i \in N$, a non-empty set of possible actions A_i, and a preference relation $\geq_i$ over the product $\times_{j \in N} A_j$ of possible actions of all players. The intuitive idea behind this definition can be put as follows. Each player i can choose any action from his alternatives A_i. If all the players have made their choice, we get what is called an 'action profile'. Intuitively, such a profile is one of the possible courses which a game may take. If our players are $1, \ldots, n$ and if they choose actions $a_1, \ldots, a_n \in \times_{j \in N} A_j$ then that's one possible 'run' of the game.

Players are assumed to choose an action which has a preferred result. Preferences over results are given by the preference relations $(\geq_i)$ which are taken to depend wholly and only on the particular actions which the players may choose. Thus, if the players $1, \ldots, n$ choose actions $a^* = a_1, \ldots, a_n$, respectively, then the result may be better for one player i than when they choose $b^* = b_1, \ldots, b_n$. In that case, we find that $a^* >_i b^*$, that is, $a^* \geq_i b^*$ and not $b^* \geq_i a^*$ Obviously, it may be the case that $a^* >_i b^*$ and $a^* >_j b^*$ for two profiles a^* and b^* and players i and j. (This is the case, typically, when two players have competing or conflicting interests.) In general it is assumed that preference relations are reflexive, transitive and complete.

It may be clear, even from these introductory comments, that the consequences of a particular choice of player i for action a_i generally depend, not only on this particular choice, but also on the choices which the other players make. Thus, if the players $1, \ldots, n$ choose the action profile $a^* =$

$a_1, \ldots, a_n$, respectively, then player a_i may be happy about the result, but if player i sticks to his choice a_i, while the others $1, \ldots, i-1, i+1, \ldots, n$ happen to choose $b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n$, the result may be less welcome for i, of course. On the other hand, if we may assume that the other players $1, \ldots, i-1, i+1, \ldots, i_n$ choose $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, respectively, then player i is assumed to choose an action a_i such that outcome or profile $a^* = a_1, \ldots, a_n$ is at least as good as any alternative profile $a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n$ which may result from an alternative choice of i for b_i. A note on notation: if we have a profile $a^* = a_1, \ldots, a_n$, then we use a^*_{-i} to indicate the list of profile's strategies of all players except i—i.e., $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$—and we use (a^*_{-i}, b_i) to indicate the profile which is like a^* with the sole difference that i chooses b_i in stead of a_i. Typically, of course, $a^* = (a^*_{-i}, a_i)$.

In order to clarify these notions a bit more, consider the following somewhat stylized example. A famous two player game is a 'coordination game' called 'Bach or Strawinsky'.[4] In this game two persons want to go out. They can choose between the performance of a concert of Bach and the performance of a concert of Strawinsky. One player (Bonnie) prefers to go to Bach, the other (Clyde) prefers Strawinsky, but the main concern of both players is to go out together. Formally, this corresponds to a game $\langle N, (A_i), (\geq_i) \rangle$, where

(3) the set of players $N = \{b, c\}$ consists of Bonnie and Clyde

(4) the set of possible actions of Bonnie and Clyde $A_b = A_c = \{B, S\}$ consist of (a choice for) Bach and Strawinsky

The profiles of this game are (B, B), (B, S), (S, B), and (S, S), where (x, y) indicates the profile which obtains when Bonnie chooses x and Clyde chooses y. Since Bonnie and Clyde definitely prefer to go out together, they both prefer (B, B) and (S, S) over the two other profiles (B, S) and (S, B). Since Bonnie moreover prefers Bach, she also prefers (B, B) over (S, S) and (B, S) over (S, B). Similarly, Clyde prefers (S, S) over (B, B), and (B, S) over (S, B). The preferences of Bonnie and Clyde, $>_b$ and $>_c$, can thus be summarized as follows:

(5) $(B, B) >_b (S, S) >_b (B, S) >_b (S, B)$

(6) $(S, S) >_c (B, B) >_c (B, S) >_c (S, B)$

A convenient representation of two player games can be given in a two-dimensional matrix, in which the various rows represent the possible actions of player one (Bonnie) and the columns the possible actions of player two (Clyde):

4. Originally known as 'The Battle of the Sexes'.

(7)

	B	S
B	$(3,2)$	$(1,1)$
S	$(0,0)$	$(2,3)$

In this matrix, we have filled in payoff pairs (n, m) which indicate the relative payoff of a specific action profile (x, y) for Bonnie and Clyde, respectively. Thus, the pair $(3, 2)$ indicates the relative payoff of Bonnie (3) and Clyde (2) when Bonnie and Clyde both choose Bach. For Bonnie this constitutes a better payoff then the one in which both choose Strawinsky, because in that case we find a relative payoff pair $(2, 3)$ where Bonnie's payoff (2) is less than 3. For a similar reason, the last profile is better for Clyde, because he prefers a joint choice for Strawinsky over a joint choice for Bach. However, both of these profiles are better than the two in which they do not go out together, and in which they at best reach a payoff of only one.

3.2 Nash Equilibria as Solutions

One of the central notions in game theory is that of a solution concept. In general, solution concepts are abstract and formal specifications of certain optimality recipes. They relate to the reasonable choices which players may make, given some notion of rationality and common knowledge. A very well known solution concept is that of a 'Nash Equilibrium'. A Nash Equilibrium of a strategic game $\langle N, (A_i), (\geq_i)\rangle$ is an action profile $a^* \in \times_{j\in N} A_j$ such that:

(8) $\forall i \in N$ and $a_i \in A_i$: $a^* \geq_i (a^*_{-i}, a_i)$

Intuitively, this says the following. 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. For no player i is there any alternative a_i for the action a^*_i which he chooses in a^*, by means of which she can get a better payoff, *given that* all the other players choose as they choose in a^*. A Nash Equilibrium clearly need not give the best possible result which one player might prefer. A player gets the best payoff *relative to* the choices of the other players in the profile, and this really is an equilibrium because this holds for all players.

If we now return to the example which we discussed above we can see that it has two Nash Equilibria, the ones in which both Bonnie and Clyde choose Bach, and the one in which both choose Strawinsky. It is expedient to see why these profiles qualify as equilibria. The profile (B, B) is a Nash Equilibrium because, *given that* Bonnie chooses Bach, the best possible outcome for Clyde obtains when he chooses Bach as well (since $(B, B) >_c (B, S)$), while *given that* Clyde chooses Bach, Bach is also the very best choice for Bonnie (since $(B, B) >_b (S, B)$). Something analogous holds of the (S,S) equilibrium. In both profiles, none of the two players has reason to deviate from the choice he actually makes. Surely, when Bonnie considers the Nash Equilibrium (S,S) she might reason as follows: "well, I better choose Bach than Strawinsky, because given that choice, it is better for Clyde to choose Bach as well, and I like (B,B) better than (S,S)" and therefore choose Bach after all. However, this type of reasoning does not by itself constitute a sound solution concept, *because if Clyde also reasons this way*, he will choose Strawinsky, and the outcome is (B,S), a profile which is worse, for both Bonnie and Clyde, than the outcome of each of the two mentioned equilibria. The nice point about the two Nash Equilibria in the Bach or Strawinsky game is that the two equilibria are not absolutely optimal profiles for both players, but optimal profiles relative to the other's choices. Both equilibria are satisfying for both players in this sense, or 'stable'.

In the definition of a Nash equilibrium, the only preferences that really count are those between two action profiles a^* and b^* if their only difference lies in the choice of i, i.e., if $a^*_{-i} = b^*_{-i}$. Furthermore, non-strict preferences, where both $a^* \geq_i b^*$ and $b^* \geq_i a^*$, do not count either. (In a Nash Equilibrium, players may have alternative options which are equally good, as long as they are not strictly better.) For this reason, Nash Equilibria in two player games can be visualized by drawing arrows between two profiles on the same row, or in the same column, with the following meaning: $\leftarrow$ means 'player 2 strictly prefers the left profile,' $\rightarrow$ means 'player 2 strictly prefers the right profile,' $\uparrow$ means 'player 1 strictly prefers the top profile,' and $\downarrow$ means 'player 1 strictly prefers the bottom profile.' The Bach or Strawinsky game then boils down to the following table:

(9)
$$\begin{array}{c|cc|} & B & S \\ \hline B & \circ \leftarrow & \\ & \uparrow & \downarrow \\ S & & \rightarrow \circ \\ \hline \end{array}$$

If in such a table no arrow leaves from a certain cell, then the corresponding profile is a Nash Equilibrium, here indicated by $\circ$. This diagram clearly shows the dependence of the two preferences of each player upon the possible choices of the other. Player 1 (Bonnie) has $\uparrow$ in case Clyde chooses Bach, and $\downarrow$ if Clyde chooses Strawinsky. Similarly, Clyde's preferences ($\leftarrow$ and $\rightarrow$) vary with the possible choices of Bonnie (Bach and Strawinsky, respectively).

3.3 Interpretation Games

From these introductory remarks the reader may already feel some connection between the notion of a solution concept and that of optimality. Both rely on a notion of 'better then' and both acknowledge a form of non-perfect optimality. Actually, we can formulate the optimality theoretic interpretation as an interpretation game.

An interpretation game is played between two players, an (abstract) speaker (S) and an (abstract) hearer (H). On the one hand, the speaker wants to communicate a certain meaning and she has to choose a suitable formulation for it; on the other, the hearer gets confronted with a certain formulation, and he has to assign it a suitable interpretation. Thus, the speaker's possible actions are given by the set of possible representations, the hearer's actions are given by the set of possible meanings, and the profiles are pairs of representations and possible meanings. Optimality theoretic preferences $<$ next can be used to define preference relations $>_S$ and $>_H$ over these pairs, and given these preference relations, some pairs of representation and meaning come out as optimal. Finally, when we evaluate for optimality, we always look along one dimension at a time. An optimal profile is one for which no player has a strictly better alternative, given that the other dimension remains fixed.

By way of illustration, consider a very simple and stylized example. Suppose that we have two names, 'Bach' and 'Strawinsky', or 'b' and 's', for short, and two possible referents, Bach (B) and Strawinsky (S). Suppose that we also have two semantic constraints, according to which 'b' preferably refers to B, and 's' to S. This game can be displayed as follows:

$$
(10)\qquad \begin{array}{c|ccc|} & B & & S \\ \hline \text{‘}b\text{’} & \circ & \leftarrow & \\ & \uparrow & & \downarrow \\ \text{‘}s\text{’} & & \rightarrow & \circ \\ \hline \end{array}
$$

Trivially, this interpretation game of the Bach or Strawinsky variety has two Nash Equilibria, which also constitute two optimal interpretations ('b',B) and ('s',S). For given that S wants to refer to B, he'd better use 'b' and given that H hears 'b', the interpretation better be B. Similarly, for the profile or interpretation ('s',S). As trivial as the example may be, it certainly shows the parallel in the type of reasoning involved in the determination of optimality as an equilibrium.

Let us now turn to two more interesting examples reminiscent of one we discussed above, viz., (1):

(11) Bill loves himself.
(12) Bill loves him.

In a matrix, the interpretation of the two sentences can be rendered as follows:

(13)
$$\begin{array}{r|ccc|} & Lbb & & Lbx \\ \cline{2-4} \textit{`self'} & \circ & \leftarrow & \\ & \uparrow & & \\ \textit{`him'} & & \rightarrow & \circ \\ \cline{2-4} \end{array}$$

There are two possible representations, 'self', which is short for (11) and 'him', short for (12). Assuming that these are evaluated in a context where Bill is salient already, there are two possible interpretations: that Bill loves himself (*Lbb*) and that Bill loves someone else (*Lbx*), a person which presumably is to be found in the context. The arrows indicate the preferences resulting from principle (B) and (DOAP):[5]

(B) If two arguments of the same semantic relation are not marked as being identical, interpret them as being distinct

(DOAP) Don't Overlook Anaphoric Possibilities

Like we said, it is assumed that (B) is stronger than (DOAP). Given this, the profile ('him',*Lbb*) is ruled out by ('self',*Lbb*) because it violates (B) and there is a better alternative, and this is indicated by $\uparrow$. Similarly, and as $\leftarrow$ indicates, ('self',*Lbx*) is ruled out by ('self',*Lbb*), because it violates (DOAP). Finally, although, ('him',*Lbx*) violates DOAP, it is better than ('him',*Lbb*), since the latter violates (B), which is judged a stronger constraint. As the pictures shows, the matrix has two Nash Equilibria, ('self',*Lbb*) and ('him',*Lbx*), precisely the two representation meaning pairs argued for.[6]

Before we carry on, it is expedient to inspect some general properties of interpretation games. It is easily seen that the following holds:

Observation 1 (Optimality Subsumes Nash)

- a profile is strongly optimal if and only if it is a Nash Equilibrium

The next observation relies on the assumption that the ordering relation $>$

5. The arrows in these matrices thus do not show the rankings of the constraints themselves, but the effects of their rankings on the preferences of the speaker and the hearer, respectively. As will be shown in more detail below, different constraints and different rankings may eventually yield the same preferences for speaker and hearer.

6. We thank Reinhart Blutner for pointing out a flaw in an earlier presentation we gave of de Hoop's analysis.

is well-founded, an assumption enforced by Jäger's requirement that it is (cf. below):

Observation 2 (No Nash, No Optimality)

- every interpretation game has a Nash Equilibrium

Proof: Given that $>$ is well-founded there is at least one (r, m) such that there is no $(r', m') < (r, m)$; a forteriori, there is no (r', m) or (r, m') such that $(r', m) >_S (r, m)$ or $(r, m') >_H (r, m)$, so (r, m) is a Nash Equilibrium. End of Proof.

The last observation is of interest from a linguistic perspective. In Game Theory, the absence of Nash Equilibria is not at all unusual, for instance in the case of zero-sum games like 'Heads or Tails', which can be displayed as follows:

(14)

	H	T
H	$(0,1)$	$(1,0)$
T	$(1,0)$	$(0,1)$

or

$$\begin{array}{c|ccc|} & H & & T \\ \hline H & & \leftarrow & \\ & \downarrow & & \uparrow \\ T & & \rightarrow & \\ \hline \end{array}$$

Well-foundedness of $>$ means that we are dealing with a particular type of game, in which solutions are guaranteed to exist. It is easily acknowledged that this makes sense: if an interpretation game were to have no solutions, then communication would be quite a void enterprise indeed.

A couple of other more general observations can be made at this point. Of course, an interpretation game would also be void if all profiles were Nash Equilibria. In that case any representation could be associated with any interpretation. With an eye on the use of language in communication, the ideal situation would obtain if the set of solutions is a one-to-one relation between the set of possible representations and the set of possible meanings. Interesting mixed cases can be characterized as well. Ambiguity obtains in situations in which the set of solutions is one-to-many; when the solutions are many-to-one we have synonymy; and when certain possible meanings do not occur in solutions we have expressive incompleteness.

4 Games and Weak Optimality

We have seen above that Blutner's strong version of two-dimensional OT can be neatly formulated using the game-theoretical concept of a Nash Equilibrium. However, Blutner (1998), and subsequently Jäger (1999) and Zeevat

(1999), have employed a 'weak' notion of optimality which is more subtle than the one we discussed in section 2. In this section we discuss this refinement, and show that it also can be given a very intuitive Game Theoretical formulation.

4.1 Blutner / Jäger Optimality

In his (1998) paper, Blutner argues that the strong notion of optimality presented in section 2 is not entirely satisfactory. This notion does not enable us to account for Horn's (1984) division of pragmatic labour, the intuition that unmarked forms tend to be used for unmarked situations and marked forms for marked situations.

To account for cases where Horn's *division of pragmatic labour* is relevant, Blutner (1998) then proposes a *weak* version of two-dimensional OT, according to which the two dimensions of optimization are mutually related:

(15) Two-dimensional OT (Weak Version) a representation-meaning pair (r, m) is *super-optimal* iff it satisfies both the Q- and the I-principle, where:
  (Q) (r, m) satisfies the Q-principle iff there is no other pair (r', m) which satisfies the I-principle such that $(r', m) > (r, m)$
  (I) (r, m) satisfies the I-principle iff there is no other pair (r, m') which satisfies the Q-principle such that $(r, m') > (r, m)$

(Notice that this definition employs *strict* preferences over representation meaning pairs.) A possibly more transparent formulation of super-optimality has been proposed by Jäger (ms):

(16) a representation-meaning pair (r, m) is *optimal* iff:
  (Q) there is no other optimal pair (r', m): $(r', m) > (r, m)$
  (I) there is no other optimal pair (r, m'): $(r, m') > (r, m)$

Under the assumption that $>$ is transitive and well-founded, Jäger observes

(17) a representation-meaning pair is optimal in the Jäger sense if and only if it is super-optimal in the Blutner sense

Jäger's assumptions about $>$ can be argued to be pretty harmless. Transitivity, of course, is a very natural property of the 'better than' relation $>$ and well-foundedness is natural, too.

The important difference between the weak and strong notions of optimality is that the weak one accepts (super)-optimal representation-meanings pairs that would not be optimal according to the strong version. It typically allows marked expressions to have an optimal interpretation, although both the expression *and* the cases they describe have a more efficient, or more typical, counterpart. Consider, for instance, the following minimal pair discussed by Horn (1984):

(18) Lee stopped the car.
(19) Lee made the car stop.

The use of *unmarked* lexical causative *stopped* in (18) has intuitively the result that the sentence will be about an event where the car stopped in the *stereotypical* way, i.e. where the driver of the car stepped on the brake pedal. This by itself can be explained by means of the strong version of two-dimensional OT, and corresponds to a Nash-Equilibrium; unmarked is preferred to marked, and stereotypical ways of stopping cars are easier to understand than alternative unusual methods. But the strong version cannot explain why also the marked form, (19), has an interpretation; the interpretation where the car was stopped in an unusual way (pulling the emergency brake, telekinesis, etc.). It is easy to see, however, that the weak version of two-dimensional OT can explain why (19) gets this interpretation. The marked form gets the atypical interpretation, because this form-meaning pair is optimal: (i) the alternative sentence (18) doesn't get this atypical interpretation, and (ii) we prefer to refer to the typical situation by using (18) instead of (19).

For another example where, because of the division of pragmatic labour, the more specialized, or more complex, form of two in principle co-extensive expressions will be associated with the less preferred reading, look at the two following sentences discussed by Levinson (1987):

(20) He_i wants $\text{PRO}_{i,j}$ to win.
(21) He_i wants $\text{him}_{i,j}$ to win.

Although a full pronoun like "him" could in principle refer to the same object as the null PRO, the selection of the full pronoun over its empty counterpart in fact signals the absence of the coreferential reading. On the assumption that coreferentiallity is the preferred, or typical, option, strong optimality can explain why (20) gets the coreferential reading. But we need weak optimality to explain why also (21) gets a reading, namely the less preferred non-coreferential one. The reason is, again, that the preferred coreferential reading is *blocked* due to the existence of the less lexicalized expression (20) that *could* have been used.

Before we turn to the game-theoretical formulation of the Blutner / Jäger notion of (weak) optimality, it is expedient to present Jäger's algorithm for computing optimal representation-meaning pairs. The algorithm computes which pairs are optimal and which are blocked, in a recursive manner. It starts off with empty sets OPT and BLO of optimal and blocked pairs and terminates when all pairs are either optimal or blocked. It is convenient to indicate the pairs which have not yet been classified as pairs which are still in the game: GAM $= \overline{\text{OPT} \cup \text{BLO}}$. (Thus, at the start of the algorithm, all pairs are in the game; in the end GAM is empty.) The algorithm is defined as follows:

(22) OPT $= \emptyset$; BLO $= \emptyset$;

while GAM $\neq \emptyset$:

OPT $=$ OPT $\cup \{(r,m) \notin$ BLO $\mid \neg\exists(r',m') \in$ GAM: $(r',m') > (r,m)\}$;

BLO $=$ BLO $\cup \{(r,m) \notin$ OPT $\mid \exists(r',m)$ or $(r,m') \in$ OPT$\}$;

return OPT;

By means of this procedure, first all the strongly optimal representation-meaning pairs are selected as OPT; then those pairs are selected as blocked for which there is an optimal alternative along the Q- or I-dimension; then those for which there is no better alternative in the game are selected as OPT, etc. When all pairs are, thus, categorized, the algorithm returns the set of Jäger optimal (i.e., Blutner super-optimal) pairs as output. In what follows, these are called *BJ-optimal*.

4.2 A Game-Theoretical Definition of BJ-Optimality

We have seen that the notion of strong optimality corresponds to that of a Nash Equilibrium. Now, although weak or BJ-optimality and Nash are also closely related, they are not the same, of course. BJ-optimality is a weaker (or softer) notion so that the set of Nash Equilibria of an interpretation game is or can be a proper subset of the optimal solutions. For instance, for some representation meaning pairs (r,m) there may be 'better' alternatives (r',m) or (r,m'), which however do not qualify as optimal, if there are yet other alternatives (r',m') which are.

A nice illustration can be given by means of a reanalysis of de Hoop's case of 'self' versus 'him', which is suggested to us by Reinhard Blutner. According to this analysis, there are two constraints at work, an expressive constraint 'referential economy' (RE) and an interpretive constraint 'local antecedent':

(RE) a reflexive element is preferable to a pronoun

(LA) a syntactic domain must contain a pronoun's antecedent

The effect of these constraints can be modeled by means of the following matrix for the corresponding interpretation game:

$$\begin{array}{ll|ccc|} & & Lbb & & Lbx \\ \cline{3-5} & \textit{‘self’} & \circ & \leftarrow & \\ (23) & & \uparrow & & \uparrow \\ & \textit{‘him’} & & \leftarrow & \\ \cline{3-5} \end{array}$$

where ‘self’ is again short for the sentence ‘Bill loves himself’ and ‘him’ for ‘Bill loves him’. In this case there is a clear preference for using sentence ‘self’ (the two $\uparrow$’s), and a preference for interpreting ‘self’ and ‘him’ as Bill (the two $\leftarrow$’s). As can be seen from the diagram, this game has only one Nash Equilibrium: (‘self’,Lbb), the only profile from which no arrow leaves. However, there is also a BJ-optimal profile (‘him’,Lbx) which is not a Nash Equilibrium. For, although there are better alternatives (‘him’,Lbb), and (‘self’,Lbx), these are themselves both overruled by the alternative (‘self’,Lbb). In other words, although, (‘him’,Lbb) $>_H$ (‘him’,Lbx), and (‘self’,Lbx) $>_S$ (‘him’,Lbx), these preference do not count because the preferred alternatives are each blocked by the Nash / optimal (‘self’,Lbb), since (‘self’,Lbb) $>_S$ (‘him’,Lbb) and (‘self’,Lbb) $>_H$ (‘self’,Lbx).

In the representation of an interpretation game we can visualize this kind of blocking by removing arrows. That is, if a profile points to a Nash Equilibrium, then all pointers to that profile can be removed. If we, thus, remove the arrows pointing to profiles which point to the equilibrium $\circ$ in the example above, then we get the following, derived game:

$$\begin{array}{ll|ccc|} & & Lbb & & Lbx \\ \cline{3-5} & \textit{‘self’} & \circ & \leftarrow & \\ (24) & & \uparrow & & \\ & \textit{‘him’} & & & \circ \\ \cline{3-5} \end{array}$$

In the resulting interpretation game we find two Nash Equilibria, corresponding to the two BJ-optimal solutions in the original game. This result can be generalized for more involved games with more than two representations and meanings. In such more involved games, the removal of preferences may yield games with new equilibria, and these in their turn may block yet other alternatives. Thus, if we successively keep on removing preferences for blocked profiles, then we collect more and more possible solutions, and if this process reaches a fixed point, then all the resulting Nash Equilibria of the fixed point correspond to the BJ-optimal pairs in the original game. As a matter of fact, such a procedure is the Interpretation Game Theoretical counterpart of Jäger’s algorithm.

Formally, this procedure can be specified as follows. Let $\mathcal{I}_0$ be an interpretation game $\langle N, (A_S, A_H), (>_{S,0}, >_{H,0})\rangle$, with $>_i$ a *strict* preference relation. Then we define the game $\mathcal{I}_{n+1}$—which is the game $\mathcal{I}_n$ with updated preferences—as follows:

(25) $\mathcal{I}_{n+1} = \langle N, (A_S, A_H), (>_{S,n+1}, >_{H,n+1})\rangle$ with
1. $>_{S,n+1} = >_{S,n} \setminus \{(y,z) \mid \exists x \in NE^{\mathcal{I}_n}\colon x >_{H,n} y\}$ and
2. $>_{H,n+1} = >_{H,n} \setminus \{(y,z) \mid \exists x \in NE^{\mathcal{I}_n}\colon x >_{S,n} y\}$

(In this definition $NE^{\mathcal{I}_n}$ indicates the set of Nash Equilibria of game $\mathcal{I}_n$.) If we now construct a sequence of interpretation games $\mathcal{I}_0, \ldots, \mathcal{I}_n, \ldots$ and if we find that $\mathcal{I}_{n+1} = \mathcal{I}_n$, then:

Observation 3 (BJ-Solutions are Nash in Updated Games)

- the BJ-optimal solutions of $\mathcal{I}_0$ are the Nash Equilibria of $\mathcal{I}_n$

This fact can be proved by comparing the update of preferences with Jäger's algorithm for computing optimal solutions. Jäger's procedure involves the iterated generation of optimal and blocked profiles. In the first run of this procedure profiles are accepted as optimal which are Nash Equilibria in $\mathcal{I}_0$[7] and next those are blocked which have an optimal alternative. It is relatively easily seen that:

1. updates of preferences preserve Nash Equilibria
2. if an update produces a new Nash Equilibrium, then the same profile was BJ-optimal at earlier stages
3. if we reach a fixed point $\mathcal{I}_n$, then all profiles either are a Nash Equilibrium (have no arrow leaving that profile), or are blocked (point at a Nash Equilibrium)

Here we witness one merit of viewing optimality theoretic interpretation in terms of (interpretation) games: BJ-optimal solutions can be characterized by means of the independently motivated and well-studied notion of a Nash Equilibrium.[8]

---

7. Since the procedure starts with empty sets of blocked and optimal profiles, the selected optimals (r, m) are those for which there is no preferred alternative (r', m'); of course it may be that there is such an alternative for a Nash Equilibrium, in case $r' \neq r$ and $m' \neq m$. However, if (r, m) really is a Nash Equilibrium, then it will never get blocked, and as soon as (r', m') is qualified as either optimal or blocked at some stage, then (r, m) gets accepted as optimal at the next stage. Wellfoundedness of Jäger's $>$ guarantees this effect.

8. The Game Theoretical formulation of BJ-optimality is close in spirit to von Neumann and Morgenstern (1944)'s notion of a Stable Set in a coalitional game. Stable Sets are minimal sets of outcomes for which there are no other preferable *stable* outcomes. Although the concept is framed in terms of outcomes of coalitional games, the idea is clearly similar. Cf., e.g., Osborne and Rubinstein (1994: p. 278*ff*) for more discussion.

The update procedure defined above can be illustrated by means of a somewhat artificial but illuminating example. Suppose the possible representations are linearly ordered, so that we can number them: $r_0, r_1, \ldots$, and that the possible meanings are linearly ordered, too: $m_0, m_1, \ldots$. In this game $\mathcal{I}_0$ there is one Nash Equilibrium, which is (r_0, m_0). If we update the preferences in this game, then all H's preferences for $(r_1, m_0), (r_2, m_0), \ldots$ are removed, because (r_0, m_0) is a better Nash Equilibrium for S, and S's preferences for $(r_0, m_1), (r_0, m_2), \ldots$ are removed because (r_0, m_0) is a better Nash Equilibrium for H. Thus, in $\mathcal{I}_1$, profile (r_1, m_1) comes out as Nash Equilibrium as well, because the preferences for (r_1, m_0) and (r_0, m_1) have been removed. But then we can update again, and remove all H's preferences for $(r_2, m_1), (r_3, m_1), \ldots$ and S's preferences for $(r_1, m_2), (r_1, m_3), \ldots$. Thus, in $\mathcal{I}_2$, profile (r_2, m_2) comes out as Nash Equilibrium as well. In short, we will find that in game $\mathcal{I}_n$ we have Nash Equilibria (r_i, m_i) for all $i \leq n$, so that we construct the diagonal as the solution of $\mathcal{I}_0$.

The last example also constitutes inspiration for the following proposition:

Observation 4 (Linearizing Unambiguous Interpretation Games)

- if the set of solutions of an interpretation game is a one-to-one relation between representations and meanings, then the preferences in the game can be equivalently stated by means of a linear order of representations and meanings

Proof. If the solutions constitute such a one-to-one relation, and if we order the solutions, then we can identify the i-st representation r_i with the representation in the i-st solution, and the i-st meaning with the meaning in the i-st solution; then we can take H's preferences to be defined by precedence in the sequence of meanings, and S's preferences by precedence in the sequence of representations, and the resulting set of solutions is the diagonal, the set of solutions we started out with. End of Proof.

4.3 On Two $\times$ Two Interpretation Games

In this section we give a systematic study of two $\times$ two interpretation games, that is games with four profiles. If we, thus, restrict our attention, we can in principle distinguish seven possible types: one in which there is no solution, one in which there is one solution, one in which there are four, one in which there are three, and three in which there are two:

(26) $\boxed{\begin{array}{cc} & \\ & \end{array}}$ $\boxed{\begin{array}{cc} \circ & \\ & \end{array}}$ $\boxed{\begin{array}{cc} \circ & \circ \\ \circ & \circ \end{array}}$ $\boxed{\begin{array}{cc} \circ & \circ \\ \circ & \end{array}}$ $\boxed{\begin{array}{cc} \circ & \circ \\ & \end{array}}$ $\boxed{\begin{array}{cc} \circ & \\ \circ & \end{array}}$ $\boxed{\begin{array}{cc} \circ & \\ & \circ \end{array}}$

(All other types are logical permutations of these types of games.) As we already observed above the first case is excluded by Jäger's well-foundedness of $>$ and the second two are void. A three-solutions game is in a sense a combination of the first two two-solutions games. The first two-solutions game models ambiguity, the second synonymy and (expressive) incompleteness, and the last is the (ideal) diagonal type.

It is interesting to note that the last type of interpretation can again be obtained in a variety of ways. All of the following matrices have the diagonal as a solution:

(27) $\boxed{\begin{array}{ccc} \circ & \leftarrow & \\ \uparrow & & \downarrow \\ & \rightarrow & \circ \end{array}}$ $\boxed{\begin{array}{ccc} \circ & \leftarrow & \\ \uparrow & & \uparrow \\ & \rightarrow & \circ \end{array}}$ $\boxed{\begin{array}{ccc} \circ & \leftarrow & \\ \uparrow & & \uparrow \\ & \leftarrow & \circ \end{array}}$

(Besides, any matrices which is a mirror of these matrices along one of the two diagonals, yields the same result as well.) In all matrices (and their mirror-images) except (the mirror-images of the) first one, one solution is not Nash, that is, in these cases the BJ-optimality of that profile is obtained by blocked preferences. This is interesting, because it shows that one and the same result can be obtained by a variety of preferences. However, this does not mean that any statement of preferences, which gives the right results, is equally good. In order to appreciate this point, consider the pair of examples discussed in Hendriks and de Hoop (to appear), under the analysis suggested by Blutner:

(1) Often when I talk to a doctor$_i$, the doctor$_{\{i,j\}}$ disagrees with him$_{\{i,j\}}$.
(28) Often when I talk to a doctor$_i$, the doctor$_{\{i,j\}}$ disagrees with himself$_{\{i,j\}}$.

A BJ-optimal interpretation of example (1) is one in which the indices on the noun phrases "the doctor" and "him" are different, so that either "the doctor" or "him" is interpreted as anaphoric upon "a doctor", not both. An optimal interpretation of example (28) is one in which both "the doctor" and "himself" are interpreted as anaphoric upon "a doctor". These results can be obtained by the joint effect of the two constraints (RE) and (LA) which we repeat here for convenience:

(RE) a reflexive element is preferable to a pronoun
(LA) a syntactic domain must contain a pronoun's antecedent

The relevant preferences are displayed in the following diagram:

$$(29)\qquad \begin{array}{r|ccc|} & (i,i) & & \{i,j\} \\ \hline \textit{`the doctor–self'} & \circ & \leftarrow & \\ & \uparrow & & \uparrow \\ \textit{`the doctor–him'} & & \leftarrow & \circ \\ \hline \end{array}$$

This is a diagram of the third diagonal type, in which ('the doctor–him',$\{i, j\}$) is a BJ-optimal solution because the (LA)-preference for ('the doctor–him',(i, i)) is blocked by the (RE)-preference of ('the doctor–self',(i, i)) over this alternative, and because the (RE)-preference for ('the doctor–self',$\{i, j\}$) is blocked by the (LA)-preference of ('the doctor–self',(i, i)) over this alternative. However, as we argued, we could have obtained the very same result if the preferences were spelled out, alternatively, as indicated by the following diagram:

$$(30)\qquad \begin{array}{r|ccc|} & (i,i) & & \{i,j\} \\ \hline \textit{`the doctor–self'} & \circ & \rightarrow & \\ & \downarrow & & \downarrow \\ \textit{`the doctor–him'} & & \rightarrow & \circ \\ \hline \end{array}$$

In this diagram, we have encoded the effect of the *converse* of the principles (RE) and (LA), and we have obtained a mirror image of the original matrix. This time the solution ('the doctor–him',$\{i, j\}$) is optimal (Nash), and the interpretation of ('the doctor–self',(i, i)) turns out BJ-optimal, but the resulting BJ-optimal pairs are the same. Does this mean that we can get away with using the converses of any two or more principles? Certainly not. This can be appreciated when we look at a more general case, where we take more possibilities ((j, j), and $\{j, k\}$) into account:

$$(31)\qquad \begin{array}{r|ccccccc|} & (i,i) & & \{i,j\} & & (j,j) & & \{j,k\} \\ \hline \text{`the doctor–self'} & \circ & \leftarrow & & \leftarrow & & \leftarrow & \\ & \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{`the doctor–him'} & & \leftarrow & \circ & \leftarrow & & \leftarrow & \\ \hline \end{array}$$

With the principles (RE) and (LA) we get the right solutions ('the doctor–self',(i, i)) and ('the doctor–him',$\{i, j\}$). If, instead, we had adopted their counterintuitive converses, the solutions would have been, incorrectly, ('the doctor–self',(j, j)) and ('the doctor–him',$\{j, k\}$). This exercise thus shows that not any way of getting certain interpretation results is fine. It also shows that one should be careful with the notion of (BJ-)optimality, or that of a solution in interpretation games. Optimal profiles can get blocked if more options get considered (and if more constraints are involved).

5 Prospects and Conclusions

In this paper we have pointed out some parallelisms between some notions studied in Optimality Theory and in Game Theory. Optimality theoretic interpretation can be modeled in terms of an interpretation game, and both Blutner's notion of (strong) optimality, as well as the Blutner / Jäger notion of (weak) optimality, can be defined as a Nash Equilibrium of the interpretation game, or of an update of it.

We have restricted ourselves, here, in two respects. Of the various types of games studied in Game Theory we have studied only one, and we have concentrated upon only one type of solution concept. The natural question that arises is if optimality theoretic interpretation would not gain if we employed other kinds of games (extensive, in stead of strategic, games; games with imperfect, rather than complete, information) and other solution concepts. In this respect we must mention Parikh (1991), who applies Game Theory to an analysis of the process of disambiguation, and who employs extensive cooperative game with partial information. It remains an open question how Parikh's approach relates to the one discussed in this paper.

Another restriction is that we have concentrated mainly on the *formal* parallelism between optimality games. However, the parallel with the work of Parikh, and the intuitions behind the Q- and I-principles, suggest that the parallelism goes deeper. Optimality crucially involves both the speaker and the hearer, conceived of as rational agents with possibly opposing preferences. An optimal interpretation of a sentence can thus be seen as the result of (hypothetical) negotiation between two players who, with their particular beliefs and desires, engage in a communication game. Here lies an interesting parallel with the approach advocated in Merin (1997). Merin construes verbal interaction as a game in which speaker and hearer have strictly opposing preferences. It would be interesting to see if this can be given an optimality style formulation. After all, also in strictly competitive games, the players' strategies are guided by the intended optimization of the results.

Authors' Address

Paul Dekker and Robert van Rooy
ILLC/University of Amsterdam
Department of Philosophy
Nieuwe Doelenstraat 15

1012 CP Amsterdam, NL
{dekker,vanrooy}@hum.uva.nl

Acknowledgements

Earlier versions of this paper have been presented at the Optimality Theoretic Semantics Meeting at the Utrecht Institute of Linguistics OTS in January 2000, and at the Ninth CSLI Workshop on Logic, Language and Computation, Stanford, May 2000. Apart form the audiences at these occasions, the first author wishes to thank the organizers of the DIP-Colloquium in Amsterdam for providing a platform for Reinhard Blutner, Petra Hendriks, Helen de Hoop, Gerhard Jäger and Henriëtte de Swart to present their views upon optimality theoretic interpretation in Amsterdam; the second author wishes to thank the above mentioned for presenting their views, and the first author for being a cooperative player. We both thank Maria Aloni and Marie Nilsenova for additional comments. The first author is financially supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (KNAW), and the second by the Dutch Organization for Scientific Research (NWO), which are gratefully acknowledged.

References

[1] Asher, Nicholas and Alex Lascarides (1993), 'Temporal Interpretation, Discourse Relations and Commonsense Entailment', *Linguistics and Philosophy* **16**: 437–493.

[2] Atlas, Jay D. and Stephen C. Levinson (1981), '*It*-Clefts, Informativeness and Logical Form', In: P. Cole (ed.), *Radical Pragmatics*, New York: AP.

[3] Boersma, Paul (1998), *Functional Phonology. Formalizing the interactions between articulary and perceptual drives*, Ph.D. thesis, University of Amsterdam.

[4] Blutner, Reinhard (1998), 'Lexical Pragmatics', *Journal of Semantics* **15**: 115–162.

[5] Blutner, Reinhard (1999), 'Some aspects of optimality in natural language interpretation', In: H. de Hoop & H. de Swart (eds.), *OTS*^{2}. *Papers on Optimality Theoretic Semantics*, Utrecht Institute of Linguistics OTS, Utrecht.

[6] Blutner, Reinhard and Gerhard Jäger, (1999), 'Competition and interpretation: The German adverbs of repetition', manuscript, Humboldt University Berlin.

[7] Gazdar, Gerald (1979), *Pragmatics*, New York: AP.

[8] Grice, Paul (1975), 'Logic and Conversation', In: P. Cole and J.L. Morgan (eds.), *Syntax and Semantics, 3: Speech Acts*, New York: AP.

[9] Hendriks. Petra and Helen de Hoop (to appear), 'Optimality theoretic semantics', to appear in *Linguistics and Philosophy.*

[10] Hoop, Helen de and Henriëtte de Swart (to appear), 'Temporal adjunct clauses in optimality theory', to appear in *Rivista di Linguistica.*

[11] Horn, Laurence R. (1984), 'Towards a new taxonomy for pragmatic inference: Q-based and R-based implicatures', In: D. Schiffrin (ed.), *Meaning, Form, and Use in Context*, Washington: Georgetown University Press, 11–42.

[12] Jäger, Gerhard (ms), *Optimal syntax and optimal semantics*, Handout for talk at DIP-colloquium, 1999.

[13] Kratzer, Angelika (1977), 'What 'Must' and 'Can' must and can mean', *Linguistics and Philosophy* **1**: 337–375.

[14] Levinson, Stephen C. (1987), 'Minimization and conventional inference', In: J. Verschueren & M. Bertucelli-Papi (eds.), *The Pragmatic Perspective*, Amsterdam: John Benjamins, 61–129.

[15] Lewis, David (1979), 'Scorekeeping in a language game', *Journal of Philosophical Logic* **8**: 339–359.

[16] Merin, Arthur (1997), 'Information, relevance, and social decisionmaking: Some principles and results of Decision-Theoretic Semantics', In: L. Moss, J. Ginzburg, M. de Rijke (eds.), *Logic, Language, and Computation. Vol. 2*, Stanford.

[17] Neumann, John von and Oskar Morgenstern (1944), *Theory of Games and Economic Behavior*, New York: John Wiley and Sons.

[18] Osborne, Martin J. and Ariel Rubinstein (1994), *A Course in Game Theory*, MIT Press, Cambridge, MA.

[19] Parikh, Prashant (1991), 'Communication and strategic inference', *Linguistics and Philosophy* **8**: 473-51.

[20] Prince, Alan and Paul Smolensky (to appear), *Optimality Theory. Constraint interaction in generative grammar*, MIT Press, Cambridge, Mass.

[21] Sandt, Rob van der (1992), 'Presupposition projection as anaphora resolution', *Journal of Semantics* **9**: 333-377.

[22] Sanford, Tony and Simon Garrod (1981), *Understanding written language*, Chichester, England, John Wiley and Sons.

[23] Sidner, Candy L. (1983), 'Focusing in the comprehension of definite anaphora', In: M. Brady and R.C. Berwick (eds.), *Computational Models of Discourse*, MIT Press, Cambridge, 267-330.

[24] Sperber, Dan and Deirdre Wilson (1986), *Relevance: Communication and Cognition*, Blackwell, Oxford.

[25] Tesar, Bruce and Paul Smolensky (to appear), 'Learnability in Optimality Theory', to appear in *Linguistic Inquiry*.

[26] Zeevat, Henk (1999), 'Explaining presupposition triggers', In: P. Dekker (ed.), *Proceedings of the Twelfth Amsterdam Colloquium*, ILLC, Amsterdam, 19-24. ESSLLI-99 workshop on Focus and Presupposition in Multi-Speaker Discourse

[27] Zipf, George Kingsley (1949), *Human behavior and the principle of least effort*, Cambridge: Addison-Wesley.