Optimization Games: An Application

Maria Aloni*

aloni@hum.uva.nl

Abstract In an optimization game (see [5]) speaker and addressee coordinate their choice of preferred syntactic form and preferred interpretation. In this article, optimization games are used to characterize the pragmatic procedure of selection of a domain of quantification.

1 Introduction

Context plays a major part in the interpretation of quantified expressions in natural language. In this article, the attention will be focused on contextual restrictions on quantification in *intensional* constructions such as questions, propositional attitude reports and epistemic sentences. It has often been observed that our interpretation of these constructions can depend on how the relevant objects are given to us (see in particular [10], and more recently [1] and [6]). The following three examples illustrate this dependence:

The first example concerns an embedded wh-question. Suppose someone has killed Donald Duck. After a careful investigation you discover that John Smith is the culprit, you say 'John Smith did it. So I know who killed Donald Duck'. Now you want to arrest him. He is attending a masked ball. You go there, but you do not know what he looks like. You say 'This person here might be the culprit, or that person there. So I do *not* know who killed Donald Duck'. Your sentence 'I know who killed Donald Duck' obtains different truth values in the two described contexts. The evaluation of the sentence seems to be dependent on the way in which the relevant individuals are specified. These can be identified by a number of methods like naming (John Smith, Bill White, and so on) or ostension (this man here, that person there, and so on). In the first context in which identification by name is assumed, the sentence is true. In the second context, in which identification by ostension is assumed, the sentence is false.

The second example expands on a well-known situation discussed by Quine in [14]. 'There is a certain man in a brown hat whom Ralph has glimpsed several times under questionable circumstances on which we need not enter here; suffice it to say that Ralph suspects he is a spy. Also there is a grey-haired man, vaguely known to Ralph as rather a pillar of the community, whom Ralph is not aware of having seen except once at the beach. Now Ralph does not know it but the men are one and the same.' ([14], p. 179.) We can tell each half of this story separately. In one half Ralph sees the man, who is called Ortcutt, in the brown hat. In the other he sees him on the beach. From the first story you can reason as follows: 'Ralph believes that the man in the brown hat is a spy. The man in the brown hat is Ortcutt. So Ralph believes of Ortcutt that he is a spy'. From the second story you can reason as follows: 'Ralph believes that the man on the beach is not a spy. The man on the beach is Ortcutt. So Ralph does *not* believe of Ortcutt that he is a spy'. Although we do not have to assume that there is any change in Ralph's belief state, it seems unproblematic to say that Ralph believes Ortcutt to be a spy and Ralph does *not*

[*] I would like to thank Johan van Benthem, Paul Dekker, Robert van Rooy, and Henk Zeevat.

believe Ortcutt to be a spy, depending on which part of the story you are taking into consideration. Our evaluation of the sentence 'Ralph believes of Ortcutt that he is a spy' depends on how Ortcutt is identified in Ralph's belief state. If Ortcutt is specified as the man in the brown hat, the sentence is true. If Ortcutt is specified as the man seen on the beach, the sentence is false.

At last consider the following situation. A butler and a gardener are sitting in a room. One is called Alfred and the other Bill. We don't know who is who. In addition, assume that the butler committed some crime. You say: 'Alfred might be innocent. Bill might be innocent. So anybody in the room might be innocent'. But you could also have said: 'The butler did it. So it is *not* true that anybody in the room might be innocent'. Again we have a sentence – 'Anybody in the room might be innocent' – which obtains different truth values depending on how the relevant individuals are identified. If they are identified by name, the sentence is true. If identification by description is assumed, the sentence is false.

The conceptual presupposition underlying classical quantified modal logic that there exists a unique favored method of trans-world identification is undermined by these natural language examples. Different methods of identification are operative in different conversational circumstances and, as the examples above illustrate, our evaluation of intensional constructions can vary relative to these methods. Dependence on identification methods is a pervasive phenomenon in natural language. In order to account for it, [1] proposes to represent identification methods in a possible world semantics by means of sets of *separated* concepts,[1] which are called *conceptual covers* and let variables range over elements of such sets. In this analysis, different domains can be selected on different occasions. Although variables always range over the same sort of individuals, these may be differently identified.

The described style of quantification is adopted in the partition theory of question (see [7]), in Hintikka's logic for propositional attitudes (see [10]) and in an intensional dynamic semantics (see [8]). The wh-question in (1a), the *de re* belief report[2] in (2a) and the quantified epistemic sentence in (3a) are analyzed as follows:

(1) a. Who killed Donald Duck?

b. $?x_n K(x_n, d)$

(2) a. Ralph believes of Ortcutt that he is a spy.

b. $\exists x_n(x_n = o \wedge Bel_r S(x_n))$

(3) a. Anyone might be innocent.

b. $\forall x_n \Diamond I(x_n)$

The representations in (1b), (2b) and (3b) receive the standard interpretation in the three above-mentioned theories with the only exception that the variable x_n is taken to range over the set of concepts contextually selected as value for n, rather than over the set of plain individuals (see [1] for a fully detailed analysis). In this way the interpretation of questions, attitude reports and epistemic sentences is made dependent on the conceptualization of the universe of discourse which is contextually operative.

The question I will explore in the present article is how the addressee arrives to select the intended domain of quantification while interpreting these intensional

[1] An individual concept is a (total) function from a set of worlds to a set of individuals. Two concepts are separated if their values never coincide.

[2] Attitude reports like 'Ralph believes that Ortcutt is a spy' are traditionally taken to be ambiguous between a *de re* reading represented in (2) and a *de dicto* reading represented as $Bel_r S(o)$, which can be paraphrased as 'Ralph would assent to the sentence 'Ortcutt is a spy''.

constructions. People use different principles to arrive at the proper interpretation of this sort of sentences in a given context. As we will see, these principles can be crucially violated and are potentially conflicting. This suggests a formulation of these contextual domain selections in the framework of Optimality Theory (see [13]).

In **Optimality Theory** (OT) conflicts between constraints are arbitrated by ranking one constraint over the other. OT has been applied in phonology, in syntax, and, recently, also in semantics and at the semantics-pragmatics interface.

According to an **OT semantics** (see [4]) the process of interpretation of natural language sentences is ruled by a number of ordered constraints. The addressee chooses from a set of possible meanings the ones which optimally satisfy these constraints. An OT interpretation theory of this sort enables an explanation of our contextual interpretation of intensional constructions in many cases, but not all. A number of potentially problematic examples of *de re* belief reports cannot be explained by such an addressee-oriented analysis. I will suggest that a proper treatment of such examples requires a **bi-dimensional OT interpretation theory** (see [2], [3] and [17]), in which also the speaker's perspective is taken into consideration. I will follow [5] and recast bi-dimensional OT interpretation processes using notions from **Game Theory**. Pragmatic domain selections are formalized in terms of *optimization games* in which speaker and interpreter coordinate their choice towards optimal form-meaning pairs.

Dependence on identification methods is just one example of the crucial role of context in natural language use. Optimality theoretic semantics supplies us with an innovative view on how structural and contextual information interact in natural language interpretation. OT semantics has however a problem: it fails to account for the dynamic multi-agent character of communication. Game theory naturally presents itself as an elegant framework for describing the coordinated actions of speaker and addressee towards optimal interpretations.

2 Optimal Theoretic Interpretation

An OT semantics is based on a set of constraints ordered according to their relative strength, which help us in deciding between different interpretations for a given syntactic structure. The addressee has a set of alternative contents for a specific expression at her disposal. The best interpretations are those elements of the set which do better on the constraints than all other alternative candidates, where candidates that have arbitrary many violations of lower ranked constraints do better than candidates that have also one violation of a higher ranked one. In OT a candidate can only be rejected if there is a better candidate available. It can never be rejected because it violates a certain constraint.

Several examples of pragmatic selection of a method of identification provide evidence in favor of competition and ranking of linguistic constraints. As we will see, when interpreting intensional sentences in a given context, people follow general pragmatic and semantic principles. They try to satisfy these principles as much as they can. But sometimes they must violate some in order to satisfy stronger ones. This section will discuss a number of these conflicting principles and present examples illustrating how these conflicts should be arbitrated.

Interpretation Constraints The first constraint I will discuss is **ANCHOR** formulated in [17]. This principle says that interpretation should be anchored to the context. ANCHOR governs the interpretation of expressions like pronouns, tenses and our domain indices, which are assigned a value either by anaphora or by deixis, and hence should find a proper antecedent in the context. As an illustration, consider again the following examples discussed in the previous section:

(4) Alfred might be innocent. Bill might be innocent. So anybody$_n$ in the room might be innocent.

(5) The butler did it. So it is *not* true that anybody$_m$ in the room might be innocent.

In (4) and (5), indices n and m find their natural antecedents in the sets {Alfred, Bill} and {the butler, the gardener} respectively, which are made salient by the explicit mention of (one of) their elements in the preceding discourse.

Now, ANCHOR says that domain indices should find proper antecedents in the context. Still assuming that the absence of a suitable salient antecedent leads to communication breakdown is quite unrealistic. Suppose you find the following sentence written on a wall.

(6) Any$_n$ grain of sand in the desert might be a diamond.

In real life communication, people would deal with such cases by accommodating one or the other antecedent for index n.[3] So, in such situations we should allow accommodation. However, accommodation should be disallowed in case a proper antecedent is already available in the context as, for instance, in the butler examples above. On the one hand, OT semantics can capture the latter intuition by assuming a principle which prohibits the addressee the addition of new material to the context, [17] calls such a principle ***ACCOMMODATE** (see also [2]). On the other hand, OT can also account for the fact that accommodation is allowed in certain circumstances by positing that ANCHOR can overrule *ACCOMMODATE (see again [17]). If no antecedent is available as in example (6), we choose to accommodate in order to satisfy ANCHOR which ranks higher than *ACCOMMODATE. If an antecedent is already present in the context as in the butler examples, ANCHOR is satisfied. Consequently *ACCOMMODATE is the critical constraint and its violation becomes crucial. So, we prefer readings which do not involve accommodation – e.g. identification by name as value for n in (4) and identification by description for m in (5). Such interaction between ANCHOR and *ACCOMMODATE also explains the domain selections in the examples of Donald Duck and Ortcutt discussed in the previous section.

Another important constraint governing the procedure of domain selection is ***SHIFT**, which expresses a general preference for interpretations which do not involve shift of domain of quantification. In the following examples, *SHIFT is overruled by the principles **CONSISTENCY**, ***TRIVIAL** and **RELEVANCE** which formulate Grice's maxims of rational conversation. CONSISTENCY expresses preference for interpretations which do not conflict with the context. *TRIVIAL forbids under-informative interpretations. RELEVANCE expresses preference for relevant interpretations.

Suppose you are attending a workshop. In front of you lies the list of names of all participants, around you are sitting the participants in flesh and blood. You do not know who is who. Consider the following sentences uttered in such a situation:

(7) Who$_n$ is who$_m$?

(8) I don't know who$_n$ is who$_m$.

[3] Accommodation mechanisms are also ruled by violable constraints like CONSISTENCY discussed below and STRENGTH formulated in [1], chapter 4.

Intuitively a proper answer to the question 'who is who' in the given context is one which specifies a mapping between the people in the room and the names in the list. An answer like 'van Benthen is van Benthem, etc.' or 'That man is that man, etc.' would be misleading in such case. The intended meaning of this question is captured in our framework by letting the two wh-expressions range over two different sets of concepts, the one representing demonstrative identification, the other identification by name. Part of the optimal theoretical explanation of the domain selection procedure in examples (7) and (8) runs as follows. In both cases we have two conflicting constraints. On the one hand, we have *SHIFT which suggests to interpret m as n. On the other, the fulfillment of *TRIVIAL and CONSISTENCY prevents this resolution in (7) and (8) respectively. Indeed, if m is n, the question in (7) is vacuous and the assertion in (8) is inconsistent. If we assume, as it has been suggested in the literature, that general rules of rational conversations are harder than *SHIFT, we have an explanation of why we switch domain of quantification while interpreting these sentences.

The following case has been inspired by an example of van Rooy in [15]. In front of Ralph stand two women. Ralph believes that the woman on the left, who is smiling, is Bea and the woman on the right, who is frowning, is Ann. As a matter of fact, exactly the opposite is the case. Bea is frowning on the right and Ann is smiling on the left. Suppose all of a sudden Ralph starts chasing the woman on the left to bring her to a mental institution. I ask you: 'Why is Ralph chasing Ann?'. You answer:

(9) Ralph believes that Ann is insane.

There are three possible ways of interpreting this sentences in the described situation: (a) an interpretation *de re*, in which Ann is identified in Ralph's state as the woman on the left; (b) an interpretation *de re*, in which Ann is identified by name; (c) the *de dicto* interpretation. All three interpretations are consistent with the background. Interpretation (a) seems to involve a violation of *SHIFT. Indeed, my question, which explicitly uses 'Ann' to identify the relevant woman, suggests identification by name as the operative method of identification. Interpretation (b) and (c) do not involve such violation. Still, intuitively, we prefer interpretation (a) for (9) in such situation. I suggest that the reason is that only under such an interpretation the sentence would be relevant. Indeed, whether the belief attribution (9) is contributing to explain for us Ralph's behaviour depends on how Ann is identified in Ralph's belief state. Whether or not Ralph believed that Ann – who is, according to him, the woman on the right– is insane does not help explaining why he is chasing the woman on the left, whereas the fact that he believes that the woman on the left is insane does contribute to an explanation. Thus, only under interpretation (a) the belief attribution constitutes a proper answer to my question and hence is relevant. This is why the addressee selects a domain containing the concept 'the woman on the left' in such a situation, although this involves a violation of *SHIFT. By assuming that RELEVANCE is harder than *SHIFT we can account for this intuition.

Consider as further application the following case where an optimal theoretic solution is proposed for a traditional problem discussed by Kaplan in [12] in relation to Quine's analysis of relational belief (see [14]). As it is easy to see, this problem emerges for all theories of *de re* belief which involve quantification over concepts rather than objects. Suppose Ralph believes there are spies, but does not believe of anyone that she is a spy. He further believes that no two spies have the same height which entails that there is a shortest spy. In such a situation, the following *de re* sentence is intuitively false.

(10) There is someone whom$_n$ Ralph believes to be a spy.

However in our analysis, we can easily find a value for n which makes the sentence true in the described situation, namely a set containing the concept *the shortest spy*. We must explain how this counter-intuitive resolution is blocked in our theory.

The belief attribution in (10) is consistent with the common ground, only if we assume such problematic value for n. But, under such an interpretation, the sentence is crucially trivialized. Thus, the only way to satisfy CONSISTENCY here would involve a violation of *TRIVIAL. This explains why the sentence is pragmatically unacceptable in such a situation. Still, the inconsistent reading is preferred over the trivial one in this case – the sentence is intuitively judged false – and the present analysis can explain this fact as follows. While the inconsistent reading only violates CONSISTENCY, the trivial reading violates *TRIVIAL, but also *SHIFT, since our description of the context – in particular the sentence 'Ralph does not believe of anyone$_m$ that she is a spy' – suggests as active a cover which does *not* contain the concept *the shortest spy*. Since CONSISTENCY and *TRIVIAL are assumed not to be ranked in any way, the violation of the lower constraint *SHIFT becomes fatal in this case.

Summarizing, we have discussed the following principles which seem to play a role in the operation of selection of a method of identification:

ANCHOR Interpretation should be anchored to the context.

CONVERSATIONAL MAXIMS

> **CONSISTENCY** avoid inconsistent interpretations;
>
> ***TRIVIAL** avoid trivial interpretations;
>
> **RELEVANCE** prefer relevant interpretations.[4]

***SHIFT** Do not shift domain of quantification.

***ACCOMMODATE** Do not accommodate.

The following is a possible ranking consistent with the phenomena analyzed above:

ANCHOR, CONVERSATIONAL MAXIMS > *SHIFT, *ACCOMMODATE

The OT analysis discussed so far enables an explanation of the process of selection of a method of identification in many cases, but not all. As an illustration, I will present the following variation on the 'shortest spy' problem I have just discussed.

Consider the following situation. Naming is the prominent method of identification and the addressee holds as common ground that: (i) Putin is the actual president of Russia; (ii) Ralph believes that Jeltsin is the actual president of Russia; (iii) Ralph would not assent to the sentence: 'Putin is bald'. Consider now the following sentence uttered in such a situation:

(11) Ralph believes that Putin is bald.

Let A stand for identification by name and B be a conceptual cover containing the concept 'the president of Russia'. The sentence has three possible interpretations in such circumstances: (a) the *de dicto* reading; (b) the *de re* reading under A; and (c) the *de re* reading under B. According to the interpretation theory I have discussed so far, interpretation (c) would be optimal in this situation. Although such

[4] See the interesting quantitative characterization of the notion of relevance proposed by van Rooy in this volume, which possibly subsumes the three principles of rational conversation formulated here. Since these principles are not assumed to be ranked in any way in the present theory, they can be reduced to a more general principle without relevant changes in our predictions. The same holds for the principles *SHIFT and *ACCOMMODATE. I will leave this issue as a subject for future study.

an interpretation violates *SHIFT and *ACCOMMODATE, the other two alternative interpretations, which satisfy these constraints, crucially violate the higher ranked CONSISTENCY, because of clause (iii) above. On the predicted optimal interpretation (c), the sentence says that (Putin is the actual president of Russia and) Ralph would assent to the sentence 'The president of Russia is bald'. Since Ralph believes that Jeltsin is the actual president of Russia, (c) also entails the *de dicto* reading of the sentence: 'Ralph believes that Jeltsin is bald'. This prediction is clearly counter-intuitive. An intuitive explanation of why reading (c) is not preferred in such a situation is that a speaker expressing such a content by means of such a sentence would not be cooperative. Indeed, in the described situation, the same content could have been conveyed in a much more efficient way by uttering the following sentence:

(12) Ralph believes that the president of Russia is bald.

The *de dicto* reading of this alternative formulation and reading (c) of (11) convey the same information in the described situation in which the information that Putin is the actual president of Russia is part of the common ground. But the former interpretation does not involve any shift of domain or accommodation. For this reason, (12) is more efficient than (11), and, therefore, the speaker, if cooperative, should have chosen it. This is Grice's principle of cooperation. A speaker has a responsibility of what the audience will make of her sentences. In cooperative exchanges, she goes through the interpretation herself and makes sure that the intended content is as easy to obtain as possible. A cooperative speaker would never have uttered (11) to convey the information that Ralph would assent to the sentence 'Jeltsin is bald'. Therefore an interpretation of (11) which conveys such information cannot be optimal in such a situation. Note, however, that such an explanation cannot be formulated in the OT interpretation theory we have considered so far, in which inputs are given by single sentences and no reference is made to alternative sentences that the speaker might have used. In order to account for these cases, we need a more complex analysis, where the optimal solution is searched on two dimensions, rather than one: the dimension of the addressee and the one of the *speaker*, and in which the two optimization procedures of the addressee and of the speaker can refer to each other and crucially constrain each other. Such an analysis is the bi-directional Optimality Theory of Reinhard Blutner. Bi-directional OT describes the coordinated choice of speaker and addressee towards optimal form-meaning pairs. In the next section, I follow Dekker and van Rooy (D&vR) in [5] and define bi-directional OT interpretation in terms of 'interpretation games' or 'optimization games'.

3 Interpretations as Games

Since Wittgenstein, the metaphor of a language game has often been used to describe our everyday conversations. This section tries to give some substance to this old insight. Notions from the well-known field of game theory are used to explain phenomena in the semantic-pragmatic interface. The presented analysis is still in its early stages. Nevertheless, it shows that the use of game theoretical concepts in linguistics is promising, although not totally unproblematic.

The central notion we will use is that of an interpretation game. An *interpretation game* I is defined as a strategic game $(N, (A_i)_{i \in N}, (\succ_i)_{i \in n})$ involving two players, Speaker and Hearer, $N = \{\mathrm{S}, \mathrm{H}\}$. The set of alternative actions for the speaker consists of a set $A_S = \{F1, F2, ...\}$ of possible *forms*, the set of alternative actions for the hearer consists of a set $A_H = \{C1, C2, ...\}$ of possible *contents*. S chooses a suitable form $F \in A_S$ for a content $C \in A_H$ to be communicated. H

chooses a suitable interpretation $C \in A_H$ for a signaled representation $F \in A_S$. Optimality theoretic preferences are used in combination with particular goal-directed preferences to define the preference relations of Speaker $\succ_S$ and Hearer $\succ_H$. The relations $\succ_S$ and $\succ_H$ should be interpreted as strict preferences and hence are taken to be transitive, anti-reflexive and anti-symmetric.

The interpretation games I will consider are crucially played in a specific context. I will therefore identify forms with utterances and interpretations with actions[5] on the specific information state which constitutes the common-ground in the circumstances of these utterances. The fulfillment of the interpretation constraints discussed in the previous section will determine the preference relation of the hearer $\succ_H$. General principles of generation, cooperativity and the particular goals of the speaker will interplay in the determination of $\succ_S$. These preference relations are highly context dependent. In the present analysis, they are sensitive to three specific aspects of the context: (a) which identification methods are salient; (b) which information is presupposed by the speaker and by the addressee; (c) the specific intentions of the speaker. The first two factors are relevant in that they determine whether or not a possible interpretation satisfies the constraints discussed above, and hence influence the preference relation of the addressee and of the speaker, if cooperative; the third factor helps in determining which content is intended by the speaker who has authority on how her utterance should be interpreted, and hence influences the preference relation of the speaker.

In this analysis, optimality is viewed as a solution concept of an interpretation game. Optimal solutions are no longer optimal interpretations of a given expression, but optimal profiles consisting of an utterance and an interpretation. We will discuss two notions of optimality: the notion of *Nash-optimality*, and *BJ-optimality*. The first notion is nothing else than the well-known solution concept of a Nash equilibrium, which has been shown to be the game-theoretic equivalent of Blutner's notion of strong optimality. BJ-optimality is the game-theoretic counterpart of Blutner's notion of weak optimality.[6] We will see how these two solution concepts can be used to account for our intuitions about the bald president example discussed above.

Bi-directional OT: Nash- and BJ-optimality In Blutner's bi-directional OT, a mechanism compares different possible interpretations C for the same syntactic expression F and another mechanism compares different possible syntactic formulations F for the same content C. A form-content pair (F,C) is then *strongly optimal* just in case C is an optimal interpretation for F according to the first mechanism and F is an optimal form for C according to the second mechanism. D&vR have shown that this notion of strong optimality can be perspicuously formalized by means of the classical solution concept of a Nash equilibrium. Blutner's strong optimal solutions are identified with Nash equilibria in an interpretation game.

Given an action profile $a \in A$ and an action $a_i \in A_i$, let $a[i : a_i]$ denote the profile which is like a, but with player i taking action a_i.

Definition 1 [Nash-optimality] Let $I = (N, (A_i)_{i \in N}, (\succ_i)_{i \in n})$ be an interpretation game. An action profile a is *Nash-optimal* in I, NASH_I(a) iff

$$\forall i \in N : \forall a_i \in A_i : \neg(a[i : a_i] \succ_i a)$$

Intuitively, in a Nash equilibrium, every player acts optimally given the other players' actions, that is, every player's action is the best response to the choices of the other players. As an illustration consider the interpretation game depicted by means of the following matrix:

[5] See [2] who also adopts dynamic updates in an OT setting.

[6] BJ-optimality is so-called after Blutner, who has introduced the notion, and Jäger, who has proposed a more transparent formulation of Blutner's notion.

	C1	*C2*
F1	(2,3)	(4,5)
F2	(3,2)	(1,1)

In such matrices, Speaker chooses the row and Hearer the column to be played and preference relations are formulated in terms of payoff functions,[7] where the payoff pair (x, y) expresses that S gets payoff x and H gets payoff y. The game depicted by this specific matrix has two Nash equilibria, namely the profiles $(F2, C1)$ and $(F1, C2)$.

In the definition of a Nash equilibrium the only strict preferences $\succ_i$ which really count are those between two profiles a and b if their only difference lies in the choice of $i \in \{S, H\}$, i.e. if $a = b[i : b_i]$ for some b_i. For this reason we can represent Nash equilibria in interpretation games by drawing arrows between two profiles on the same row or in the same column, with the following meaning: $\rightarrow$ (or $\leftarrow$) means 'H strictly prefers the right (or left) profile', and $\downarrow$ (or $\uparrow$) means 'S strictly prefers the bottom (or top) profile'. The game above is then represented by the following table in which the Nash equilibria are immediately visualized by $\circ$:

$$\begin{array}{c|ccc|}
 & C1 & & C2 \\
\hline
F1 & & \rightarrow & \circ \\
 & \downarrow & & \uparrow \\
F2 & \circ & \leftarrow & \\
\hline
\end{array}$$

If no arrow is leaving from a profile a, then a is a Nash equilibrium. This means that a profile (F, C) is Nash-optimal in I iff for all contents $CN \in A_H$ and forms $FN \in A_S$ in I:

(i) $(F, CN) \not\succ_H (F, C)$

(ii) $(FN, C) \not\succ_S (F, C)$

By means of the notion of Nash-optimality, we can characterize anomalous interpretations. A pair (F, C) is anomalous with respect to I iff it is *not* Nash-optimal in I, and this is the case iff either C is not an optimal interpretation for F in I (clause (i) is not satisfied) or, if C is an optimal interpretation, then C could have been expressed more efficiently by an alternative form (clause (ii) is not satisfied).

The strong version of optimality characterized by the notion of a Nash equilibrium is useful to explain many standard cases, but it has been shown not to be always satisfactory. In [2], Blutner illustrates this by means of the following example due to Horn:

(13) Black Bart killed the sheriff.

(14) Black Bart caused the sheriff to die.

The lexical causative ***kill*** tends to be restricted to stereotypical causative situations (e.g. Black Bart shot the sheriff), and the marked construction in (14) tends to refer to more marked situations (e.g. Black Bart caused the sheriff's gun to backfire by stuffing it with cotton). The general tendency illustrated by this example seems to be that 'unmarked forms tend to be used for unmarked situations and marked forms for marked situations' ([11], p. 26). This tendency has been called by Horn ***the division of pragmatic labour.***

This case can be formalized by means of the following interpretation game:

[7] Preference relations can be expressed in terms of payoff functions $(u_i)_{i \in N}$, where $u_i : A \rightarrow R$ is the payoff function of player i. Action profiles with higher payoff are preferred.

$$\begin{tikzcd} & C1 & C2 \\ F1 & {} \arrow[r] \arrow[d] & {} \arrow[d] \\ F2 & {} \arrow[r] & \circ \end{tikzcd}$$

where $F1$ and $F2$ stand for the marked and the unmarked forms respectively and $C1$ and $C2$ stand for the the marked and the unmarked situation respectively. By the notion of Nash-optimality we can account for the fact that (13) picks up stereotypical situations. Unmarked forms ($F2$) are preferred over marked forms ($F1$), and stereotypical situations ($C2$) are easier to understand than atypical situations ($C1$). The profile unmarked form-unmarked situation ($F2, C2$) is Nash in such a game. But Nash-optimality is not sufficient to explain why (14) obtains the unusual interpretation. Indeed, no interpretation is selected for the marked form $F1$. The profile marked form-marked content is intuitively chosen because (i) the alternative unmarked form does not get the marked interpretation and (ii) we prefer to use the unmarked form to express the unmarked situation. Now, by means of the notion of optimality defined in terms of a Nash equilibrium we cannot capture this kind of reasoning. A profile is Nash-optimal iff it is optimal for Speaker and optimal for Hearer and these two checks for optimality are independent of each other. The search for the optimal choice for one player is not influenced by the preference relation of the other player. In order to account for H's reasoning in this case, we need a notion in which the two optimization procedures of the hearer and of the speaker can refer to each other and constrain each other. Such a notion is Blutner's notion of ***weak optimality.*** BJ-optimality is the perspicuous game-theoretical formulation of such notion (see [5] for further discussion).

Definition 2 [BJ-Optimality] Let $I = (N, (A_i)_{i \in N}, (\succ_i)_{i \in N})$ be an interpretation game. Then the set BJ_I of BJ-optimal solutions in I is defined as follows:

$$BJ_I = NASH_{I_n}$$

where I_n is the fixed point, i.e. $I_{n+1} = I_n$, of the sequence of games $I_0, \ldots, I_m, \ldots$ constructed as follows:

(i) $I_0 = I$

(ii) $I_{n+1} = (N, (A)_{i \in N}, (\succ_{i_{n+1}})_{i \in N})$ with

(a) $\succ_{S_{n+1}} = \succ_{S_n} \setminus \{(y, z) \mid \exists x \in NASH_{I_n} : x \succ_{H_n} y\}$;

(b) $\succ_{H_{n+1}} = \succ_{H_n} \setminus \{(y, z) \mid \exists x \in NASH_{I_n} : x \succ_{S_n} y\}$.

In the construction of I_{n+1} you eliminate preferences for profiles y which are blocked in I_n. A profile y is blocked in a game, if there is a Nash-optimal profile x which is preferred to y in that game. If $I_{n+1} = I_n$, then the Nash equilibria of I_n are the BJ-optimal solutions in I_0. That is, if an action profile a is a Nash-optimal solution in the fixed point game of the sequence generated from a game I, then a is BJ-optimal in I.

The intuitive idea of this construction is that Nash-optimal profiles block less preferred ones and preferences for blocked profiles are overruled. As an illustration, let us go back to the game I determined by Horn's sheriff example. The sequence generated from such game consists of the two games represented in the following matrices where blocked profiles are indicated by $\perp$:

$$I_0: \quad \begin{tikzcd} & C1 & C2 \\ F1 & {} \arrow[r] \arrow[d] & \perp \arrow[d] \\ F2 & \perp \arrow[r] & \circ \end{tikzcd} \qquad\qquad I_1: \quad \begin{tikzcd} & C1 & C2 \\ F1 & \circ & \perp \arrow[d] \\ F2 & \perp \arrow[r] & \circ \end{tikzcd}$$

I_0 is I, and I_1 is obtained from I_0 by eliminating preferences for the two blocked profiles $(F2, C1)$ and $(F1, C2)$. I_1 has two Nash-optimal solutions: $(F1, C1)$ and $(F2, C2)$. Since no preference can be eliminated in the next step of our construction (i.e. $I_1 = I_2$), these two profiles are the two BJ-optimal solutions of I.

$$\begin{tikzcd}
 & C1 & C2 \\
F1 & \mathrm{BJ} \arrow[r] \arrow[d] & \bot \arrow[d] \\
F2 & \bot \arrow[r] & \circ
\end{tikzcd}$$

Profile $(F1, C1)$ is BJ-optimal because, although two arrows depart from it, these preferences do not count since they are blocked by the Nash-optimal $(F2, C2)$. In the notion of BJ-optimality, a player's perspective on optimization is crucially constrained by the other player's perspective and *vice versa*. We can now capture H's intuitive reasoning in the sheriff case. The profile marked form-marked content $(F1, C1)$ is intuitively chosen because (i) the alternative unmarked form $F2$ does not get the marked interpretation $C1$ ($(F2, C1)$ is blocked) and (ii) we prefer to use the unmarked form $F2$ to express the unmarked situation $C2$ ($(F1, C2)$ is also blocked). The hearer chooses the marked $C1$ rather than the unmarked $C2$ as interpretation for $F1$, because she can reason as follows: if the speaker had wanted to communicate $C2$, she would have chosen the Nash-optimal $F2$.

Application Let us see now how the bald president case discussed in the previous section can be accounted for by means of these optimization games. For ease of reference, I restate the situation. Naming is the prominent identification method. The common ground contains the following information: (i) Putin is the actual president of Russia; (ii) Ralph believes that Jeltsin is the actual president of Russia; (iii) Ralph would not assent to the sentence: 'Putin is bald'. In such context, Speaker says the following sentence:

(15) Ralph believes that Putin is bald.

Intuitively, a rational addressee H can do two things in such a situation: either refute to perform any action or consider revising her state with the information that Ralph would assent to the sentence: 'Putin is bald'. In any case, H does not update with the information that Ralph would assent to the sentence: 'Jeltsin is bald'. This last action was predicted as optimal by the one-dimensional OT interpretation theory I introduced in the previous section. Let us see whether the two-dimensional theory I have just described does any better here.

I propose to characterize such a situation by means of the following interpretation game:

$$\begin{tikzcd}
 & C1 & C2 \\
F1 & {} \arrow[r] & {} \arrow[d] \\
F2 & {} \arrow[u] \arrow[r] & {}
\end{tikzcd}$$

$F1$ and $F2$ are the utterances of the sentences (15) and (16) respectively:

(16) Ralph believes that the president of Russia is bald.

For ease of reference, I will denote the first action by '*Putin*' and the second action by '*the president*'.

Let us see now how $C1$ and $C2$ are characterized. Let A be naming and B be a set containing the concept 'the actual president of Russia'. Assume A and B are the only two domains available in our situation. Each of the two sentences above has then three possible interpretations.

(17) Ralph believes that Putin is bald.

a. *de dicto*: $Bel_r B(p)$

b. *de re* under A: $\exists x_A[x_A = p \wedge Bel_r B(x_A)]$

c. *de re* under B: $\exists x_B[x_B = p \wedge Bel_r B(x_B)]$

(18) Ralph believes that the president of Russia is bald.

d. *de dicto*: $Bel_r B(r)$

e. *de re* under A: $\exists x_A[x_A = r \wedge Bel_r B(x_A)]$

f. *de re* under B: $\exists x_B[x_B = r \wedge Bel_r B(x_B)]$

Given our characterization of the situation, these six possible interpretations collapse in only two different possible actions on the relevant common ground. I write $UP_\sigma(\phi)$ to denote the set of the potential outcomes of updating a state σ with ϕ. Let σ stand for the common ground in the described situation. In a standard dynamic semantics, we then obtain the following equivalences:[8]

(α) $UP_\sigma(a) = UP_\sigma(b) = UP_\sigma(e)$

(β) $UP_\sigma(c) = UP_\sigma(d) = UP_\sigma(f)$

The six interpretations above collapse in only two possible alternative actions for H on the relevant state σ: the action in (α) which consists in eliminating those possibilities in σ in which it is true that Ralph would assent to the sentence: 'Putin is bald'; the action in (β) which consists in eliminating those possibilities in σ in which it is true that Ralph would assent to the sentence: 'The president of Russia is bald'. For ease of reference, I will denote the first action by $UP(put)$ and the second action by $UP(pres)$. I will identify $C1$ and $C2$ with these two possible updates.

Let us turn now to the preference relations $\succ_H$ and $\succ_S$. Hearer's preferences are obtained by the following OT analyses for the two relevant utterances and contents (in the diagrams, violations are indicated by (*), and fatal violations by !(*)):

F1	CONS	*SHIFT, *ACC
C1	!(*)	
C2		(*) (*)

F2	CONS	*SHIFT, *ACC
C1	!(*)	
C2		

Interpretation $C2$ is optimal for both syntactic inputs, because action $C1$ on the assumed common ground leads to the absurd state and hence fatally violates CONSISTENCY. Therefore, in our game, H strictly prefers a profile $(F, C2)$ over $(F, C1)$ for all $F \in A_S$. Consistent interpretations are preferred by H over inconsistent interpretations.

The assumed arrows for Speaker are justified by the fact that '*Putin*' and '*the president*' are clearly the most efficient and cooperative way of conveying the information brought about by $UP(put)$ and $UP(pres)$ respectively. Efficient formulations are preferred by a cooperative S over non-efficient formulations.

Our game has one Nash equilibrium, namely profile ('*the president*', $UP(pres)$).

$$\begin{tikzcd}
 & UP(put) & UP(pres) \\
\textit{`Putin'} & {} \arrow[r] & {} \arrow[d] \\
\textit{`the president'} & {} \arrow[u] \arrow[r] & \circ
\end{tikzcd}$$

[8] Recall that σ supports the information that Putin is the actual president of Russia, $r = p$.

Since consistent interpretations are preferred over inconsistent interpretations and efficient formulations are preferred over non-efficient formulations, Nash-optimality selects $UP(pres)$ as optimal interpretation for '*the president*'. But Nash-optimality does not select any interpretation for '*Putin*'. Its interpretation is left open because (i) profile ('*Putin*', $UP(put)$) is anomalous since $UP(put)$ is not an optimal interpretation (it leads to inconsistency) and (ii) ('*Putin*', $UP(pres)$) is anomalous since although $UP(pres)$ is an optimal interpretation, it could have been expressed more efficiently by an alternative form. In order to account for the fact that the inconsistent interpretation of (11) under the prominent cover is preferred by the addressee over an interpretation under the problematic conceptualization we need the weaker notion of BJ-optimality. Intuitively H chooses action $UP(put)$, which would lead her to the absurd state, rather than $UP(pres)$ as a response to '*Putin*' because she can reason as follows: If Speaker had wanted to convey the consistent interpretation $UP(pres)$, then S should have chosen the more efficient formulation '*the president*'. But S chose '*Putin*'. Thus S must have meant to convey $UP(put)$. This is precisely the kind of reasoning captured by the notion of BJ-optimality. Indeed, profile ('*Putin*', $UP(put)$) is BJ-optimal in our game. Profile ('*Putin*', $UP(pres)$) is not, because overruled by the Nash-optimal ('*the president*', $UP(pres)$).

$$\begin{tikzcd}
 & UP(put) & UP(pres) \\
\textit{`Putin'} & \mathrm{BJ} \arrow[r] & \perp \arrow[d] \\
\textit{`the president'} & \perp \arrow[r] \arrow[u] & \circ
\end{tikzcd}$$

We can now explain the addressee's behaviour in our presidential example. When interpreting an utterance of 'Ralph believes that Putin is bald', she does not select a domain containing the concept 'the president of Russia' (profile ('*Putin*', $UP(pres)$) is blocked), but she rather assumes the prominent identification by name (profile ('*Putin*', $UP(put)$) is BJ-optimal). The latter action leads her to the absurd state. She can protest or she can decide to start a process of revision of her information.

4 Conclusion and Further Research

In the article I have used optimization games to describe the pragmatics of a selection of a method of identification. This analysis has allowed us to shed some light on a series of traditional difficulties emerging from the interaction between modal concepts, quantifiers and the notion of identity. Dependence on identification methods is just one example of the crucial role of contextual information in natural language use. Optimality theory and game theory have appeared as promising approaches for the explanation of this important aspect of linguistic interpretation. A number of open questions remain though in connection to the OT theory discussed in the first part, for instance, concerning the choice of the constraints and their ranking. As for the game theoretical part, the field is new and most of the theoretical questions are still unsettled. Already the characterization of the basic ingredients of an interpretation game is open to discussion. An urgent open question is how linguistic principles and particular goals should combine in the determination of the preference relations. Furthermore, in the described optimization games, speaker and interpreter have been assumed to have accurate and complete information about their payoff matrices. This is a clear limitation of the present analysis. In everyday conversations, the protagonists often fail to know the presupposition and intentions of each other and therefore can lack full knowledge of their opponent's or even their own rank of preferences. To account for such cases an extension to games is needed in which the preference relations are not common ground (see for instance the Bayesian games introduced in [9] and discussed by van Rooy in this volume). This and other issues ask for further investigation, which, however, must be left to another occasion.

References

[1] Maria Aloni. *Quantification under Conceptual Covers.* PhD thesis, University of Amsterdam, Amsterdam, 2000.

[2] Reinhard Blutner. Some aspects of optimality in natural language interpretation. To appear in *Journal of Semantics.*

[3] Reinhard Blutner and Gerhard Jäger. Competition and interpretation: the german adverbs of repetition. In *Extended Conference Abstracts: Approaching the Grammar of Adjuncts.* University of Oslo, 1999.

[4] Helen de Hoop and Petra Hendriks. Optimality theoretic semantics. *Linguistics and Philosophy,* 24(1), 2001.

[5] Paul Dekker and Robert van Rooy. Bi-directional optimality theory: An application of game theory. To appear in *Journal of Semantics.*

[6] Jelle Gerbrandy. Identity in epistemic semantics. In L. Cavedon et al., editors, *Logic, Language and Computation, Vol. III.* CSLI, Stanford, CA, 2000.

[7] Jeroen Groenendijk and Martin Stokhof. Questions. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language.* Elsevier, Amsterdam, 1997.

[8] Jeroen Groenendijk, Martin Stokhof, and Frank Veltman. Coreference and modality. In S. Lappin, editor, *Handbook of Contemporary Semantic Theory.* Blackwell, Oxford, 1996.

[9] John Harsanyi. Games with incomplete information played by 'Bayesian' players, parts I, II, and III. *Management Science,* 14:159–182, 320–334, 486–502, 1967/68.

[10] Jaakko Hintikka. Semantics for propositional attitudes. In Davis, Hockney, and Wilson, editors, *Philosophical Logic.* Reidel, Dordrecht, 1969.

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

[12] David Kaplan. Quantifying in. In D. Davidson and J. Hintikka, editors, *Words and Objections: Essays on the Work of Quine.* Reidel, Dordrecht, 1969.

[13] Alan Prince and Paul Smolensky. Optimality: from neural network to universal grammar. *Science,* 275:1604–1610, 1997.

[14] Willard V. Quine. Quantifiers and propositional attitudes. *Journal of Philosophy,* 53:101–111, 1956.

[15] Robert van Rooy. Decision problems in pragmatics. In M. Poesio and D. Traum, editors, *Proceedings of Götalog 2000.* Göteborg University, 2000.

[16] Robert van Rooy. Relevance of communicative acts. this volume.

[17] Henk Zeevat. The asymmetry of optimality theoretic syntax and semantics. To appear in *Journal of Semantics.*