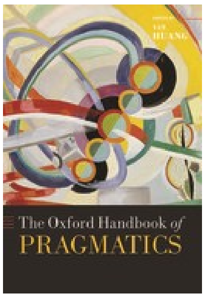


The Oxford Handbook of Pragmatics

Yan Huang (ed.)

https://doi.org/10.1093/oxfordhb/9780199697960.001.0001

Published: 2013 **Online ISBN:** 9780191749858 **Print ISBN:** 9780199697960

CHAPTER

5 Formal Pragmatics

Reinhard K. Blutner

https://doi.org/10.1093/oxfordhb/9780199697960.013.002 Pages 101–119

Published: 05 September 2013

Abstract

In this article, three theoretic frameworks are discussed: optimality-theoretic, game-theoretic, and decision-theoretic pragmatics, the last being based on Ducrot's argumentation theory. The close similarities between optimality-theoretic and game-theoretic pragmatics are pointed out. Concerning decision-theoretic pragmatics, some arguments are provided demonstrating that an independent, argumentation-theoretic grounding is neither needed nor useful. Rather, it seems more appropriate to incorporate the argumentation-theoretic insights into a general Gricean-oriented theory of natural language interpretation, let it be optimality-theoretic pragmatics or a game-theoretic variant.

Keywords: argumentation theory, bidirectional optimization, blocking, bounded rationality, decision theory, evolutionary game theory, fossilization, optimal interpretation, optimality theory, quantum probabilities

Subject: Pragmatics, Linguistics

Series: Oxford Handbooks

5.1 Introduction

AS prominently demonstrated in theoretical physics, the formal language of mathematics may be very useful for describing aspects of reality. That does not mean that the mathematical instruments are intended to capture a precise picture of reality. Instead, processes of abstraction and idealization are omnipresent (Stokhof and van Lambalgen 2011), generating an apparently very close fit between pre-existing, flawless mathematical structures and an idealized/abstracted picture of reality that is studied in science.

In the field of natural language pragmatics, some researchers working on relevance theory (RT) or optimality-theoretic (OT) pragmatics take a similar naturalistic stance and claim that basic principles of cognitive psychology can be applied for grounding the basic mechanisms of natural language interpretation. The view of placing natural language pragmatics within the scope of a naturalistic (explanatory) approach is not without problems. This has to do with the normative character that is attributed to the Gricean setting. Speakers, as Grice (1975: 45) put it, must 'make their contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which (they) are engaged'.

If a person acts in a particular situation in a particular way we can ask *why* she did it the way she did; alternatively, we can ask if what the person did was *reasonable*, and if other options were possibly more reasonable in the given situation. Good Griceans are expected to ask the second type of questions (to take the normative stance), whereas the first question is expected to be asked by cognitive scientists (typically taking a naturalistic stance). Obviously, the Gricean principle of cooperation as stated above is normative, and so are Grice's conversational maxims.

Even though the normative and the naturalistic aspects of understanding human actions can be clearly separated from each other, in most cases it does not follow that they predict different action patterns. The idea of a rational world is not so irrational ↳ to be excluded from ordinary affairs. Evolutionary game theory has presented us with many examples demonstrating that the reasonable is naturally arising (Axelrod 1984). In other words, though there is a philosophical gap between Gricean pragmatics as a normative theory and pragmatic frameworks such as relevance theory as a scientific, explanatory theory of natural language, there is not a deep empirical conflict between an interpretation-oriented pragmatics and a speaker ethics. It seems that the speaker would do better to be cooperative or pretend to be cooperative if they want to use language to bring about effects on hearers. p. 102

In the present article, I will discuss three recent approaches which are crucially based on formal mathematical instruments to perform their analyses, sometimes being strictly naturalistic, sometimes being based on the normative stance. The theoretic frameworks I will discuss are optimality-theoretic pragmatics, game-theoretic pragmatics, and decision-theoretic pragmatics. I will highlight both the similarities and the essential differences between these frameworks.

5.2 Optimality-Theoretic Pragmatics

Optimality theory is an integrated approach to cognition that combines the advantages of symbolic, constraint-based models with the advantages of subsymbolic, neuron-style models of cognition (cf. Smolensky and Legendre 2006). In the study of natural language, OT was successfully applied to the main linguistic disciplines including phonology, morphology, and syntax, and also to the explanation of natural language acquisition and other performance traits. OT pragmatics is an application of the integrated approach to the domain of Gricean pragmatics. It has its origin in the attempt to explain certain phenomena of lexical pragmatics (Blutner 1998) and is inspired by the optimal interpretation approach proposed by Hendriks and de Hoop (2001).

The view of seeing OT pragmatics within the scope of a naturalistic (explanatory) approach to cognition (as represented by the main proponents of OT) brings it close to relevance theory (RT) which likewise takes the naturalistic stance (Sperber and Wilson 1986/1995). There is another point of agreement that brings OT pragmatics and RT closely together. This point concerns the fact that both theories view the division of labour between semantics and pragmatics in a similar way. Both follow the tradition of radical pragmatics and accept these three claims (e.g. Jaszczolt 2010):

1. There is a level of logical form or semantic representation. The representations of this level do not necessarily provide truth conditions. Rather, they underspecify truth-conditional content in a number of ways.

2. There is a mechanism of enriching underspecified representations; sometimes this mechanism is called development of logical form. The result of this p. 103 ↳ development is propositional content. It expresses the utterance meaning of the expression under discussion.

3. There is a level of implicatures proper, understood as separate thoughts implied by the utterance. It is implicit propositional content that can be inferred from the explicit content mentioned in 2.

Obviously, the consensus is about rejecting the Gricean doctrine of literal meaning (logical form conforms to literal meaning), accepting the role of underspecification (logical forms are underspecified with regard to the expressed semantic content), and acknowledging that implicature is a graded category (some implicatures are closer to logical form than others).

5.2.1 Three Variations on Grice

In this subsection I will discuss three variants of Gricean pragmatics: (i) RT (Sperber and Wilson 1986/1995), (ii) Levinson's (2000) theory of *presumptive meaning*, (iii) the Neo-Gricean approach (Atlas and Levinson 1981; Horn 1984; Huang 2009; see also Huang 2007). I will show how OT can formalize these three approaches and systematically relate them. In this connection it is useful to introduce the distinction between *global* and *local* approaches to conversational implicatures (cf. Chierchia 2004). According to the global (neo-Gricean) view one first computes the (plain) meaning of the sentences; then, taking into account the relevant alternatives, one strengthens that meaning by adding in the implicature This contrasts with the local view, which first introduces pragmatic assumptions locally and then projects them upwards in a strictly compositional way where certain filter conditions apply. Representatives of the global view are Gazdar (1979), Atlas and Levinson (1981), Soames (1982), (Horn 1984), Krifka (1995), Blutner (1998), Sauerland (2004), Sæbø (2004), and Geurts (2010); the local view is taken by RT (Sperber and Wilson 1986/1995, Carston 2002), Levinson (2000), and Chierchia (2004).

Usually, the globalists argue against the local view and the localists against the global view. I will argue, instead, that proper variants of both views are justified if a different status is assigned to the two views: global theories provide the standards of rational discourse and correspond to a diachronic, evolutionary scenario; local theories account for the shape of actual, online processing, including the peculiarities of incremental interpretation. In this way, I will argue that seemingly conflicting approaches such as relevance theory and the neo-Gricean approach are much more closely related than the adherents of one side or the other might expect. OT will prove its power of unification in giving hints on how to relate these different frameworks in a systematic way.

RT assumes the representational/computational view of the mind, and, on this basis, gives a naturalization of pragmatics adopting Jerry Fodor's language of thought p. 104 ↳ hypothesis (Fodor 1975). The central thesis of RT is the *communicative principle of relevance*, according to which utterances convey a presumption of their own optimal relevance. In other words, any given utterance can be presumed:

- to be at least relevant enough to warrant the addressee's processing effort;

- to be the most relevant one compatible with the speaker's current state of knowledge and her personal preferences and goals.

From these two assumptions relevance theorists derive the following general procedure that the cognitive system follows in comprehending an utterance (cf. Sperber, Cara, and Girotto 1995: 95): (a) test possible interpretations in their order of accessibility, and (b) stop once the expectation of (optimal) relevance is satisfied (i.e. a certain context-dependent threshold value of relevance is reached). The procedure makes sure that the wanted effect (a certain value of relevance) is reached with the minimal cognitive effort.

Levinson's (2000) theory of *presumptive meaning* is a chameleon that in a certain sense adapts general assumptions of RT and in another sense crucially conflicts with RT, for instance in assuming more than one basic principle (*maxim*) for formulating the interpretational mechanism. In short, these are the general assumptions:

- Differing from both RT and the standard neo-Gricean view, Levinson assumes *three* levels of meaning corresponding to sentence(-type) meaning, utterance-type meaning, and utterance-token meaning
- utterance-type meanings are in correspondence with Grice's generalized conversational implicatures. They are a matter of preferred interpretation calculated by a particular default mechanism. Basically, there are three such defaults or heuristics:
  - Q-heuristic: What isn't said is not the case
  - I-heuristic: What is expressed simply is stereotypically exemplified
  - M-heuristic: What's said in an abnormal way isn't normal
- In contrast to Grice's generalized conversational implicatures, which are calculated in a global manner, presumptive meanings are local, i.e. they arise at the point at which they are triggered (for instance, the word *some* triggers the default interpretation NOT ALL via the Q-heuristic). The feature of local pragmatics is essential to artificial intelligence pragmatics (e.g. Hobbs and Martin 1987) and likewise to RT.

Presumptive meanings are very useful for understanding natural language interpretation, especially for explaining the predominantly incremental character of utterance comprehension.

Neo-Griceans (Atlas and Levinson 1981; Horn 1984, 2005a; Blutner 1998; Atlas 2005; Huang 2009) are
p. 105 assuming two countervailing optimization principles: the ↳ Q-principle and the R-principle.[1] The first is oriented to the interests of the hearer and looks for optimal interpretations; the second is oriented to the interests of the speaker and looks for expressive optimization. Here is a standard presentation of the two principles (cf. Horn 1984, 1989, 2004, 2005a):

The Q-Principle (hearer-based)

Make your contribution sufficient!

Say as much as you can! (modulo R)

(Grice's first quantity maxim and the first two manner maxims)

The R-Principle (speaker-based)

Make your contribution necessary!

Say not more than you must! (modulo Q)

(Grice's second quantity maxim, relation maxim and the second two manner maxims)

It is tempting to identify the Q-principle with Levinson's Q-heuristic and the R-principle with the I-heuristic. However, they are not identical though there is a correspondence between them. The difference has to do with the different status of *principles* in the global, neo-Gricean pragmatics on the one hand and *heuristics* (*defaults*) in Levinson's local pragmatics on the other hand. According to the neo-Gricean picture the principles constitute a kind of communication game—either between real speakers and hearers or between fictive speakers and hearers in the mind of a language user. In this game both principles are applied in a recursive way (corresponding to the modulo clause in the formulation of the principles). In Levinson's theory, no such interaction between real or fictive speakers/hearers takes place. Instead, presumptive meanings are default interpretations and they are processed in a nearly automatic way. No 'mind-reading' facilities or other mechanisms of controlled processing are required.[2] The difference will become quite clear in the following subsection when I give formalization in terms of bidirectional OT.

5.2.2 Bidirectional OT

Bidirectional optimality theory falls within the family of linguistic models that are based on the p. 106 optimization of linguistic output against a system of ranked constraints ↳ (Blutner 2000; Blutner and Zeevat 2004; Blutner, de Hoop, and Hendriks 2005; Benz and Mattausch 2011). This theory provides a general procedure of optimization of the relation between form and meaning, simultaneously optimizing in both directions, from meaning to form, and from form to meaning. This distinguishes bidirectional optimality theory from unidirectional optimality-theoretic semantics (Hendriks and de Hoop 2001)—optimizing from form to meaning—and from unidirectional optimality-theoretic syntax (Grimshaw 1997)—optimizing from meaning to form.

To put it in a nutshell, bidirectional optimality theory evaluates form–meaning pairs. As described in Blutner (2000), there are two ways of defining optimality in a bidirectional setting, a strong way and a weak way. The strong version is based on the standard definition of optimality, applying this to candidate pairs instead of output elements.

In the following I will define an OT system for a set F of forms and a set M of meanings as a pair $\langle \mathbf{Gen}, \succ \rangle$ consisting of a generator $\mathbf{Gen} \subseteq \mathbf{F} \times \mathbf{M}$ that gives us the set of all potential form–meaning pairs and an ordering on elements of $\succ$ **Gen**. Informally, the relational statement $f' \succ_m f$ says that the pair $\langle f', m \rangle$ satisfies the system of (ranked) constraints better than the pair $\langle f, m \rangle$; the statement $m' \succ_f m$ says that the pair $\langle f, m' \rangle$ satisfies the system of (ranked) constraints better than the pair $\langle f, m \rangle$ (borrowing the notation used by Franke 2009). In the strong version of bidirectional OT, a form–interpretation pair $\langle f, m \rangle \in \mathbf{Gen}$ is considered to be (strongly) optimal iff

- Interpretive Optimization: there is no pair $\langle f, m' \rangle \in \mathbf{Gen}$ such that $m' \succ_f m$
- Expressive Optimization: there is no pair $\langle f', m \rangle \in \mathbf{Gen}$ such that $f' \succ_m f$.

Informally, the first clause says that *m* is the optimal interpretation of *f*, and the second clause says that *f* is an optimal expression for *m*.

The weak version of bidirectional optimality is less restrictive than the strong one and normally allows for more solution pairs. The original formulation (Blutner 2000) is close to the (recursive) formulation of Horn's Q- and R-principle and it allows us to derive Horn's division of pragmatic labour (Horn 1984, 1989, 2004, 2005a)—i.e., the general propensity that 'unmarked forms tend to be used for unmarked situations and marked forms for marked situations' (Horn 1984: 26). In the following I adopt Jäger's reformulation of the original definition (Jäger 2002). A form–interpretation pair $\langle f, m \rangle \in \mathbf{Gen}$ is considered to be superoptimal (or weakly optimal) iff

- Interpretive Optimization: there is no superoptimal pair ⟨*f, m′*⟩ ∈ **Gen** such that $m' \succ f\ m$
- Expressive Optimization: there is no superoptimal pair ⟨*f ′, m*⟩ ∈ **Gen** such that $f' \succ m\ f$.

This formulation looks like a circular definition, but Jäger (2002) has shown that this is a sound recursive p. 107 definition under very general conditions (well-foundedness of the ↳ ordering relation). This recursive definition is our expression of the communication game constituted by the neo-Gricean picture as described at the end of section 5.2.1.

A simple example should illustrate the difference between the two optimization concepts. The example I will use goes back to McCawley (1978) who observed that the distribution of productive causatives (in English, Japanese, German, and other languages) is restricted by the existence of a corresponding lexical causative. Whereas lexical causatives such as in (1a) tend to be restricted in their distribution to the stereotypical causative situation (direct, unmediated causation through physical action), productive (periphrastic) causatives as in (1b) tend to pick up more marked situations of mediated, indirect causation. For example, (1b) could have been used appropriately when Black Bart caused the sheriff's gun to backfire by stuffing it with cotton.

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

The example presents a scenario with two forms, *kill* and *cause to die*, and two interpretations, *dir* and *indir*, referring to direct (stereotypic) causation and indirect causation, respectively. Assuming that the semantics for *kill* and *cause to die* admits the same range of interpretations, we get four form–meaning pairs described by **Gen**: ⟨*kill, dir*⟩ , ⟨*kill, indir*⟩ , ⟨*cause to die, dir*⟩ , ⟨*cause to die, indir*⟩ . Table 5.1 shows these four pairs together with two markedness constraints, called F and M. The constraint F (for forms) marks complex forms; in the present case it marks the *cause to die* construction. The other constraint is M (for meanings) and it marks the complex interpretations; in the present case it marks the *indirect* interpretation. The effect of the F-constraints results in the relation *kill* $>_x$ *cause to die* for any interpretation *x*, and the effect of the M-constraint results in the relation *dir* $>_y$ *indir* for any form *y*.

The left part (a) of Table 5.1 illustrates the strong version of bidirectionality. Since for both forms, the direct interpretation gives the optimal interpretation and for both interpretations, the form *kill* gives the optimal expression, the pair ⟨*kill, dir*⟩ , is the only strongly optimal pair (marked with the symbol ✌ in Table 5.1a). As a consequence, the form *cause to die* is blocked in each potential interpretation.[3] Unfortunately, the prediction of total blocking is intuitively not correct in the present example; instead, the blocking is partial —allowing ⟨*cause to die, indir*⟩ as a second solution pair as predicted by Horn's division of pragmatic labour.

Table 5.1b shows that the *weak* version of bidirectionality can explain the effects of partial blocking without the stipulation of extra constraints that link forms and meanings directly; in particular, it can explain why the marked form *cause to die* gets the marked interpretation *indir*. This is a consequence of the *recursion*[4] p. 108 implemented ↳ in weak bidirectionality: the pairs ⟨*kill, indir*⟩ and ⟨*cause to die, dir*⟩ are not superoptimal. Hence, they cannot block the pair ⟨*cause to die, indir*⟩ , and it comes out as a new superoptimal pair (likewise marked with the symbol ✌ in the table). In this way, the weak version accounts for Horn's pattern of *the division of pragmatic labour*. In the literature several algorithms have been proposed for formulating explicit recursive mechanisms for calculating superoptimal pairs (Jäger 2002; Beaver and Lee 2004; Franke and Jäger 2012).

Table 5.1 Strong and weak bidirectionality using markedness constraints: (a) shows strong bidirectionality, (b) shows weak bidirectionality (superoptimality)

(a)

		F	M
✌	⟨*kill, dir*⟩		
	⟨*kill, indir*⟩		*
	⟨*cause to die, dir*⟩		*
	⟨*cause to die, indir*⟩	*	*

(b)

		F	M
✌	⟨*kill, dir*⟩		
	⟨*kill, indir*⟩		*
	⟨*cause to die, dir*⟩	*	
✌	⟨*cause to die, indir*⟩	*	*

Unfortunately, there are some doubts about the psychological reality of such mechanisms as models of online natural language interpretation (e.g. Blutner 2010). Instead, it has been proposed to take the diachronic perspective into account, as clearly expressed by Horn (1984). Hence, in the framework of optimality-theoretic pragmatics it is very natural to take weak bidirectionality as expressing a basic principle of natural language change. As a consequence, bidirectional optimization has nothing to do with online processes that run during normal language interpretation/production. Rather, the results of bidirectional optimization are routinized or fossilized—a phenomenon that takes place on an evolutionary timescale. According to this evolutionary view of bidirectionality, form–meaning pairs that have been determined by bidirectional optimization constitute fixed relations to a learner who sets out to acquire the p. 109 language. ↳ No learner, indeed no user of the language, needs to perform a bidirectional computation for any form–meaning pair she encounters.

Table 5.2 Strong and weak bidirectionality using linking constraints

		F→M	*F→*M	>>	F→*M	F*→M
✌	⟨*kill, dir*⟩				*	
	⟨*kill, indir*⟩	*				
	⟨*cause to die, dir*⟩		*			
✌	⟨*cause to die, indir*⟩					*

Let us come back to our simple example in order to get an idea of what ‘fossilization’ could mean. Rather than considering markedness constraints, Table 5.2 presents so-called linking constraints that connect the form level with the interpretational level. In the present example there are precisely four independent linking constraints. The linking constraint F→M says that simple (unmarked) forms express simple interpretations. Hence, this is a straightforward formalization of Levinson’s (2000) **I**-heuristic as an OT constraint. The constraint *F→*M says that complex forms express complex interpretations, and this is an expression of Levinson’s **M**-heuristic.[5] The two remaining linking constraints express the opposite restrictions. In the present case linking constraints can be seen as lexical stipulations that fix a form–interpretation relation in an instance-based way. Assuming that a general learning mechanism ensures that the two latter linking constraints are finally ranked lower than the former two, then the result of strong bidirectional optimization is the same as the result of weak bidirectional optimization discussed before. In addition, it can be seen from Table 5.2 that unidirectional optimization (taking the hearer’s or the speaker’s perspective) is sufficient already and gives exactly the same results.

It is not the place here to discuss real candidate mechanisms for the fossilization process. Such processes can be best understood when related to an offline mechanism that is based on bidirectional learning (Blutner, Borra, Lentz, Uijlings, and Zevenhuijzen 2002; Benz 2003, 2006; Jäger 2004; van Rooij 2004c). In these approaches the solution concept of weak bidirectionality is considered as a principle describing the results of language change: superoptimal pairs emerge over time in language change. This confirms the age-old theory that synchronic structure is significantly informed by diachronic forces.

p. 110 Let me come back now to the earlier goal of giving an OT reconstruction of the three variations on Grice. For reconstructing Levinson’s (2000) presumptive meaning theory, unidirectional optimization is sufficient where a system of OT constraints has to be formulated conforming to his I, Q, and M heuristics and Levinson’s putative ranking Q > M > I. The unidirectional optimization procedure (interpretive optimization) is to conform with a local approach to conversational implicatures, one which satisfies the requirements of incremental interpretation.

The neo-Gricean approach, on the other hand, is globalist in nature. Hence, the idea of (weak) bidirectional optimization fits best with this theory and can be used for a straightforward formalization. Unsurprisingly, this conception can be seen best from a diachronic perspective, as long as we take a naturalistic stance towards Gricean pragmatics. As a model of actual language interpretation (or production) this approach does not make real sense and was never designed for this purpose.

Like Levinson’s (2000) approach, RT conforms to the localist approach and can be formulated in terms of unidirectional optimization. Let us stipulate a constraint EFFECT for describing the wanted effect (a certain value of relevance) and a constraint EFFORT for describing the cognitive effort. Then the stipulation EFFECT > EFFORT makes sure that the wanted effect is reached with the minimal cognitive effort. Obviously, there are many questions left concerning the concrete content of the constraints EFFECT and EFFORT, and the RT literature contains a number of specifications. These specifications typically have the character of linking constraints. It might be interesting to investigate recent OT models of pragmatics (see section 5.4) in the light of the general structure of RT—a task that goes beyond what can be done in the present paper.

I have mentioned already that there is a relation between diachronic and synchronic systems, and I have introduced the term *fossilization* for describing the relevant transfer. Given the existence of this transfer, it can be demonstrated that the three variations on Grice discussed here are much more closely related than the occasional polemics led us to expect.

Bidirectional OT has been used for describing a series of phenomena and observation in the domain of natural language pragmatics. In the following I give an overview of some of these applications without going into any technical or empirical details.

- *Disambiguation*. Gärtner (2004a, b) analyses Icelandic object shift and differential marking of (in)definites in Tagalog, addressing the issue of disambiguation and partial iconicity in natural language.

- *Binding theory*. Mattausch (2004a, b) introduces the influential work of Levinson on the origin and typology of binding theory (summarized in Levinson 2000; see also Huang 1994/2007, 2000a) and reformulates the different historical stages assumed by Levinson in bidirectional optimality theory. Mattausch's work is of essential importance as one of the first in-depth studies showing the importance of the diachronic view for bidirectional OT.

- p. 111 ↳ *Discourse particles and presupposition*. Zeevat (2002, 2004) treats discourse particles within an extended OT reconstruction of presupposition theory. In another paper, Zeevat (2007) provides a full solution to the projection problem for presuppositions.

- *Complex implicatures*. Blutner (2007) gives an OT account of implicature projection and explains the relevant theoretic distinction between implicatures and explicatures in terms of a neo-Gricean framework.

- *Interpretation of stress and focus*. Several articles deal with a bidirectional perspective for stress on anaphoric pronouns and the interpretation of focus (Beaver 2004; de Hoop 2004; Hendriks 2004; Aloni, Butler, and Hindsill 2007).

- *Marking and Interpretation of negation*. Henriëtte de Swart (2004) provides a bidirectional OT approach to the syntax and pragmatics of negation and negative indefinites (see also de Swart 2010).

- *Permission sentences*. A series of other articles deals with the interpretations of permission sentences and the analysis of the particular conditions which constitute a so-called *free choice* interpretation (Sæbø 2004; Aloni 2005a, 2005b; Blutner 2006).

- *Aspectual interpretation of the Dutch past tenses*. Van Hout (2007) applied bidirectional reasoning about tense forms and their aspectual meanings.

- *Lexical pragmatics*: Lexical Pragmatics investigates the processes by which linguistically specified ('literal') word meanings are modified in use. Prototypical applications include the pragmatics of dimensional adjectives (Blutner and Solstad 2000), the analysis of Dutch *om/rond* (Zwarts 2006), the pragmatics of negated antonyms (Blutner 2004; Krifka 2007), gender opposition of animate nouns (Zwarts, Hogeweg, Lestrade, and Malchukov 2009), and several examples of semantic change (Eckardt 2002).

- *Language acquisition and learning*: There are several studies that test the role of weak bidirectionality in developing interpretation and production preferences in connection with (in)definite NPs (de Hoop and Kramer 2005/2006; van Hout, Harrigan, and de Villiers 2010) and pronominal anaphors (Hendriks and Spenader 2005/2006; Hendriks, van Rijn, and Valkenier 2007; Mattausch and Gülzow 2007; Hendriks, de Hoop, Krämer, Swart, and Zwarts 2010; van Rij, van Rijn, and Hendriks 2010).

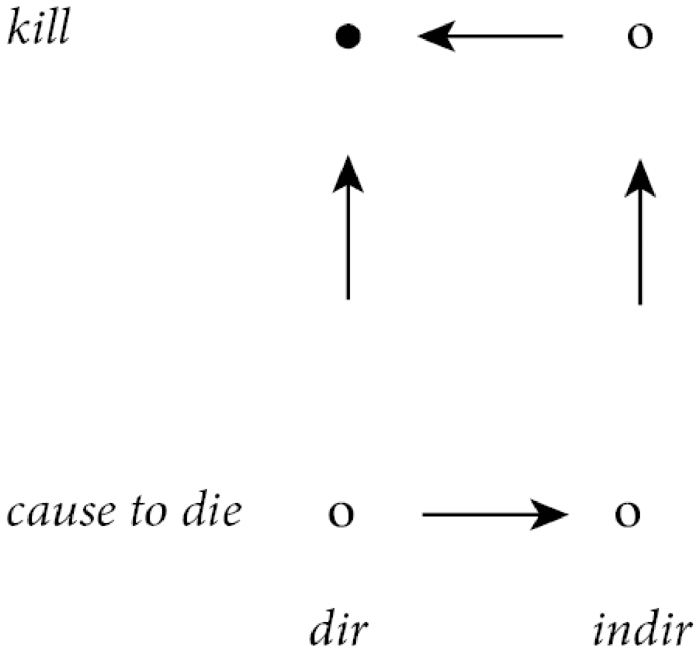
5.3 Game-Theoretic Pragmatics

Wittgenstein is widely acknowledged as the founding father who connected games with natural language interpretation. ‘In the 1950s, the later Wittgenstein famously moved away from the crystalline logical structure of the *Tractatus* to a paradigm of rule-generating “language games”’ (van Benthem 2008: 198). p. 112 Another prominent ↳ researcher applying game theory to natural language interpretation is Hintikka, who introduced evaluation games in order to specify the truth-conditional semantics of certain fragments of natural language. In 1969, David Lewis’s book *Convention* was published, which introduced the important idea of signalling games to the domain of language (Lewis 1969). These games refer to the phenomena of ‘cultural evolution’ (Hurford 1998; Steels 1998) and aim to give a mathematical justification of the formation of stable meanings (Nash equilibriums of signalling games).

In this section, I will discuss the close connection between OT pragmatics and game-theoretic pragmatics. In the literature, we find two different kinds of games capturing the strong and the weak solution concept in bidirectional OT: strategic games (Dekker and van Rooij 2000) and signalling games (Benz, Jäger, and van Rooij 2005a; Franke 2009; Franke and Jäger 2010). The interesting point is that the game-theoretic approach provides a richer system of solution concepts than the optimality-theoretic one. Further, it proposes an impressive spectrum of possible ways to approximate the intended solutions—as defined by a certain solution concept—by means of iterative reasoning protocols (Franke and Jäger 2012). Hence, we can expect from the game-theoretic approach not only a sound explication of optimality-theoretic solution concepts, but also new solution concepts and algorithms that challenge the standard optimality-theoretic ones.

Strategic games are games where players move simultaneously. This contrasts with sequential games where players move in sequence. Let us assume a strategic game with two players (called speaker and hearer). Speaker’s possible actions are given by the set of possible forms, *kill* and *cause to die* in the present example; the hearer’s possible actions are given by the set of possible meanings, *dir* and *indir* in our example. Generally, in (two-person) games, pairs of the two players’ actions are called profiles. Hence, in our example profiles are the four possible form–meaning pairs indicated by the four small circles in Figure 5.1.

Figure 5.1



Nash equilibrium (indicated by •) in a concrete example

In game theory, solution concepts are formal specifications of certain optimality concepts. They are rational/normative concepts relating to the reasonable choices which players may make. A famous solution concept is that of a 'Nash equilibrium'. A Nash equilibrium for a strategic game (with two players S and H) is p. 113 an action profile $\langle a_S, a_H \rangle$, such that each player's action is an optimal response to the choices of the other ↳ players in that profile, i.e. for the speaker there is no action x_S such that $\langle x_S, a_H \rangle \succ_S \langle a_S, a_H \rangle$ and for the hearer there is no action y_H such that $\langle a_S, y_H \rangle \succ_H \langle a_S, a_H \rangle$. In Figure 5.1 there exists exactly one Nash equilibrium indicated by the black circle. The horizontal arrows indicate the strict preferences for the hearer (the arrow directs to the stronger pair) and the vertical arrows show the strict preferences for the speaker. Informally, a pair is a Nash equilibrium if no arrow leads away from this pair.

Franke (2009) pointed out that it is conceptually not very plausible to use strategic games for modelling language use. Literally taken, the two players in a strategic game (say speaker and hearer) make their choices independently from each other. The speaker chooses the preferred (lightest) form and the hearer chooses the preferred (simplest) interpretation. Even when the resulting form–meaning pair is realized in our natural language system, the underlying solution concepts for strategic games are far from providing a plausible (causal) argument why the selected form is connected with the selected meaning.

Franke (2009) argued that OT systems would better be translated into some kind of sequential game with imperfect information where for a given meaning the speaker chooses a corresponding form to express this meaning, and the hearer subsequently tries to guess at this meaning on the basis of the uttered form. 'A natural idea is to consult signaling games, a class of games which are widely used for the study of strategic communication not just in linguistics, but also in biology, economics, and the philosophy of language (c.f. Lewis 1969, Spence 1973, Grafen 1990)' (Franke and Jäger 2012: 5).

This is not the place to give a detailed introduction into the idea of signalling games and its use in formal pragmatics as done in some recent monographs (Lewis 1969; Parikh 2001; Benz, Jäger, and van Rooij 2006; Franke 2009). Instead, I will develop a simplified basic picture and outline some relations to OT pragmatics. My exposition will follow the presentation of Jäger (2007a). The basic idea of a signalling game is rather simple. In the most straightforward case we have two players called S (the speaker) and H (the hearer). The game begins with a randomly chosen meaning *m* that is presented to S (but not to H). Next, the speaker is requested to choose a signal *f* that is transmitted to H. On the basis of this signal, the hearer H is asked to guess the meaning of *f*. If the guess is correct, i.e. H selects the meaning *m*, both S and H score one point, if not both get nothing. It is easy to see that in signalling games the interests of the two players completely coincide. Further, these games are asymmetric games because the two roles of speaker and hearer are not interchangeable.

What are the possible strategies of the two players? In the simplest case of deterministic (pure) strategies, a possible strategy of the speaker is a function from meanings to forms and a possible strategy of the hearer is a function from forms to meanings. I will write S(*m*) for applying a speaker strategy to a meaning *m* and H(*f*) for applying a hearer strategy to a form *f*. Further, a similarity function sim(*m*, *m*') is used which gives the value 1 if *m* and *m'* completely agree and gives the value 0 if they are maximally different.[6] With these prerequisites at hand, the utility function of the game is p. 114 ↳ given by the following equation assuming the prior probabilities P(*m*) are common knowledge between the two players:

$$(2)\quad \mathrm{u}(\mathrm{S},\mathrm{H}) \;= \sum\nolimits_{m} \mathrm{P}(m)\ \mathrm{sim}(m,\ \mathrm{H}(\mathrm{S}(m)))$$

There is no distinction made between the utility of the speaker and the hearer. Because the games in question are real partnership games, both players always obtain the same utility. Equation (2) expresses our basic intuition that communication is successful if the players of the signalling game understand each other. At the moment, the costs for interpreting the signal and generating the signal are ignored. However, it is not difficult to subtract a corresponding cost value c(*f*, *m*) from the left-hand side of equation (2) in order to take the costs for signalling and interpretation into account. This can be done in correspondence to a given OT system by respecting the following relations (e.g. Franke 2009; Franke and Jäger 2012):

$$(3)\quad \langle m_1,\ f_1\rangle > \langle m_2,\ f_2\rangle \quad \text{iff}\,\mathrm{c}(m_1,\ f_1) < \mathrm{c}(m_2,\ f_2).$$

Intuitively, in signalling games the two players try to find strategies that maximize the value of the utility function u(S,H). There are different ways in which this could be realized.

First consider the evolutionary interpretation of game theory (cf. Jäger 2007a). In this case the utility of a strategy is to be interpreted as the expected number of offspring of a player adopting this strategy. Technically, this is described by so-called replicator dynamics—a deterministic continuous time dynamics for sufficiently large populations. Since it is extremely difficult to solve the differential equations determining the replicator dynamics in an analytic way, Maynard-Smith (1982) developed a way to characterize the qualitative behaviour of the replicator dynamics. The central conception is that of an evolutionary stable system. Informally, an evolutionarily stable strategy is characterized by the configuration of a population that is stable in the sense that the population does not leave its state due to its inherent dynamics, and is protected against small amounts of mutation. Under certain conditions it can be proven that Nash equilibriums are evolutionarily stable. As a consequence, the superoptimal solution pairs discussed above come out as evolutionarily stable (Benz 2003; van Rooij 2004c; Benz 2006; Lentz and Blutner 2009; Franke and Jäger 2012).

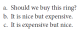
A second mechanism is iterated learning theory where the utility function (2), or a modification of it, is optimized by an iterated learning process where speaker and hearer learn from each other (cf. Kirby and Hurford 1997, 2002; Jäger 2004; Mattausch 2004a, b; Benz 2006; Mattausch and Gülzow 2007). Using straightforward OT learning models, the results are similar to those found in the evolutionary interpretation of game theory. Similar results are found by investigating reinforcement learning for signalling games (Franke and Jäger 2010).

A third idea is realized by the iterated best response model as proposed by Matsui (1992). It was recently p. 115 applied to pragmatics (Jäger 2007b; Franke 2009; Franke and ↳ Jäger 2012). This model proposes a particular evolutionary interpretation of signalling games. Other than in Darwinian evolution where new strategies only emerge due to undirected random mutation, the present model suggests that whenever a new member enters the population they may freely choose their strategy. If it is assumed that the new members are rational enough to maximize their expected utility they will choose a strategy that is an optimal response to the average strategy of the population. By repeating the addition of new members indefinitely, an optimal response dynamics is defined which is different from the standard replicator dynamics sketched above. Interesting differences include the emergence of scalar implicatures and total blocking (Jäger 2007b; Franke 2009; Franke and Jäger 2012).

In section 5.1 we discussed the normative and the naturalistic aspects of understanding human actions. Interestingly, the three approaches to signalling games discussed in this section clearly exhibit the naturalist stance. And they clearly relate to offline aspects of natural language processing (cultural evolution, language change, bidirectional learning). This idea corresponds to the understanding of weak bidirectionality which relates best to an offline mechanism that is based on bidirectional learning (Blutner, Borra, Lentz, Uijlings, and Zevenhuijzen 2002; Benz 2003; van Rooij 2004c). It suggests that the borderline between semantics and pragmatics is transparent in at least one direction: tendencies predicted from pragmatics (conversational implicatures modelled by weak bidirectionality) may become frozen or fossilized in the semantic component of knowledge representation. The details of the fossilization process are an open problem. Obviously, evolutionary game theory and variants of it may be a powerful instrument to explore different hypotheses concerning the self-organizing dynamics of language as an observationally learned and culturally transmitted communication system.

5.4 Decision-Theoretic Pragmatics

Decision-theoretic pragmatics (Merin 1999) is closely related to argumentation theory (Ducrot 1972, 1973, 1980, 1984). According to this view, utterances are normally used as premises and conclusions in arguments. It is this argumentative use in language that determines the meaning of utterances in discourse. Interestingly, this conception of meaning goes far beyond what is normally described as the truth-conditional conception of meaning. For instance, utterances with the same informational content can be used as arguments for quite different things. A famous example is due to Anscombre and Ducrot (1983):

(4) 

a. Should we buy this ring?
b. It is nice but expensive.
c. It is expensive but nice.

Assuming that the informational (= truth-conditional) content of (4b) and (4c) is the same, Anscombre and p. 116 Ducrot (1983) claim that (4b) and (4c) argue for opposite ↳ things when seen in the context of (4a): (4b) argues for not buying the ring; (4c) argues for buying it. Examples like this led many authors to believe that a purely truth-conditional semantics is not sufficient for an adequate meaning description and that the ‘argumentative potential’ of an utterance forms an essential part of its meaning.

Merin (1999) has started the formalization of the key ingredients for an argumentative theory of pragmasemantics. His theory is based on the classical conception of probability and elements of decision theory. He gives a precise definition of concepts like informativeness and relevance, and he makes precise the idea of issue-based communication. According to Merin (1999), argumentation is a probabilistic relation over epistemic states. A proposition *A* is a positive argument for a hypothesis *H* iff accepting *A* increases the probability of *H*. It is a negative argument for *H* iff accepting *A* lowers the probability of *H*. The formal expression of the argumentation relation is the 'relation of relevance':

(5) $\mathrm{r}_H(A) = \log(\mathrm{P}(A|H) - \log(\mathrm{P}(A|\neg H)$

Positive relevance means $\mathrm{r}_H(A) > 0$; negative relevance means $\mathrm{r}_H(A) < 0$. A simple consequence of the definition in (5) is that $\mathrm{r}_H(A) > 0$ iff $\mathrm{P}(H|A) > \mathrm{P}(H)$ and $\mathrm{r}_H(A) < 0$ iff $\mathrm{P}(H|A) < \mathrm{P}(H)$. For a proof one has simply to make use of the Bayesian formula.

One of the cornerstones of argumentation theory is a semantic analysis of 'but'. We will look at it since it involves an appealing application of the conception of relevance. What is the main phenomenon we have to describe? In a seminal paper, Lakoff (1971) distinguished two different uses of *but*, the 'contrast use' and the 'denial of expectation use'.

(6) a. John is tall but Sue is short.
b. John is a Republican but he is honest.

Examples like (6a) illustrate the *contrast* use of 'but'. Such examples are always symmetric, i.e. if the order of the conjuncts is reversed no significant meaning changes are induced. Further, the substitution of 'but' by 'and' does not induce a significant change of meaning. Examples like (6b) illustrate the *denial of expectation* use. Such utterances are typically not symmetric and the substitution of 'but' by 'and' leads to significant changes of meaning: (6b) suggest that Republicans are normally not honest, whereas the reverse of the conjuncts suggests that honest persons are normally not Republicans.

It is not difficult to see that the argumentative approach works pretty well for the *denial of expectation* use and allows us to express an important constraint stated by Anscombre and Ducrot (1983). As shown in Winterstein (2011), this constraint can be formulated using Merin's (1999) notion of relevance:

(7) For an utterance of the form *p but q*, there must be an *H* such that:

a. $\mathrm{r}_H(p) > 0$ and $\mathrm{r}_H(q) < 0$ (or equivalently $\mathrm{r}\neg_H(q) > 0$)

b. $\mathrm{r}\neg_H(q) > \mathrm{r}_H(p)$

p. 117 Unfortunately, a description of the contrastive case is less obvious. For example, in (6a) it is not really clear what the debated hypothesis should be.[7]

There are different kinds of criticism concerning the argumentative framework in general and Merin's decision-theoretic treatment in particular. For example, van Rooij (2004a) has argued against Merin's view that the two participants of a dialogue play a zero sum game with opposite preferences (if one agent prefers H to be true the other prefers it to be false). Instead, van Rooij argues that the participants of a dialogue are cooperative and this should be reflected in the conceptual grounding. Without going into details, some groups of examples should be mentioned where Merin's approach is not very explanatory and comes into considerable trouble: numerals, temperature expressions, disjunctions, and particularized scalar implicatures (for the details, see van Rooij 2004b, c).

In an important paper, Iten (2000) has argued that several insights of the argumentative approach could be integrated into a Gricean framework. Comparing Anscombre and Ducrot's (1977) treatment of 'but' and a recent relevance-theoretic analysis (Blakemore 2002) Iten comes to the conclusion that the two analyses 'are remarkably similar and, arguably, the grounds for choosing between them lie more with their theoretical underpinnings than with the details of the particular accounts' (Iten 2000: 665).

Iten (2000) lists and criticizes several central conjectures of argumentation theory. Among them are the following assumptions:

- Argumentation theory clearly takes the normative stance. It is a non-cognitive theory.
- The semantics of utterances provided by argumentation theory is not truth-conditional. It aims to specify the 'argumentative potential' of the utterance. Further, the argumentative potential of an utterance does not depend on the recovery of some prior truth-conditional meaning component. This contrasts with Gricean conversational implicatures depending on the recovery of 'what is said', i.e. the literal meaning that is expressed in a truth-conditional way.
- Anscombre and Ducrot (1983) use the term 'pragmatique intégrée' (integrated pragmatics) in order to indicate a uniform approach that is directed to the analysis of the non-truth-conditional aspects of utterance meanings. This term suggests that there is no semantics/pragmatics distinction in argumentation theory.
- Anscombre and Ducrot's (1983) concept of 'comparative argumentative strength' encounters counterexamples.

Concerning the lexical entries of 'but', an important consequence of these assumptions is that 'but' has to p. 118 be ambiguous since there is only one representational level ↳ where the pragmasemantics of 'but' can be described, and the different uses of 'but' are to be assumed as its different meanings.

Contrasting with the argumentative approach, Umbach (2005) does not accept any ambiguity for 'but'; rather, she stipulates a core meaning for 'but'. This core meaning comes close to what was described as the *contrast* use of 'but'. Importantly, Umbach is able to show that the *denial of expectation* use can be derived via a general mechanism of contextual enrichment. A similar treatment was proposed by Sæbø (2003), who analyses the content of the derived material as a presupposition rather than an implicature proper. Winterstein (2009) provides a critical discussion of both approaches and argues that that there are examples such as (8) that they cannot handle:

(8) Lemmy plays the bass, but Richie plays it too.

Similar arguments are put forward by Zeevat (2011). However, these counterexamples do include some systematicity that connects the different uses of 'but'. This makes it obvious that the underlying systematicity has to be described in a quite different way.

What could such an alternative analysis look like? One possibility is to make use of the idea of underspecification that is prominently connected with the view of radical pragmatics (see section 5.2) and the idea of a mechanism of pragmatic enrichment. I fully agree with Zeevat (2011: 15) who states 'that progress in the understanding of "but" is in being more precise about how to find the missing object: the question in Umbach's (2005) account, the issue that is argued for and against in the argumentative tradition, the manifest inference or the statement under objection—it is fairly easy to go from one to the other'. Assuming that Zeevat (2011) and Winterstein (2011) are on the right track paves the way for an integration of ideas of radical pragmatics with argumentation theory. Concerning formal pragmatics, it could be useful then to combine the ideas put forward in sections 5.2 and 5.3 with crucial ideas of argumentation theory.

5.5 Conclusions

In this article I have outlined the close relations between optimality-theoretic pragmatics and game-theoretic pragmatics. It has been suggested that the OT approach can profit from evolutionary game theory and variants of it in exploring different hypotheses of the self-organizing dynamics of natural language as an observationally learned and culturally transmitted communication system. Further, it has been claimed that it may be useful for the game-theoretic paradigm to overcome some bounds of the normative stance and to consider the realization of the evolutionary account and the implementation of the iterated best-response model within a plausible cognitive setting.

Finally, I have critically discussed argumentation theory and decision-theoretic pragmatics. My criticism does not imply that the argumentative framework is obsolete and not worth a serious study. There is no doubt that a lot of excellent analytical work ↳ has been done within the argumentation-theoretic framework based on interesting and new observations. Further, argumentation theory has highlighted the non-truth-conditional aspects of meaning and has made clear that some words (such as *good*, *interesting*, and *lovely*) are intrinsically subjective. However, some arguments were put forward suggesting that an integration of argumentative tradition with a (neo/post)-Gricean perspective is possible and useful.

Acknowledgement

I am indebted to Stefan Blutner, Michael Franke, Jason Mattausch, Carla Umbach, Grégoire Winterstein, and Henk Zeevat for discussion, and to Paul Dekker and Yan Huang for encouragement.

Notes

[1] In OT, these ‘principles’ correspond to different directions of optimization where the *content* of the optimization procedure is expressed by particular OT constraints. This will be pointed out in more detail in the following section.

[2] However, presumptive meanings can demand a lot of effort as soon ‘conflicts’ arise and the corresponding assumption has to be cancelled. Conflict resolution can be very resource-demanding. Hence, for the overall mechanism we have to take into account the peculiarities of controlled processing. Of course, this does not refer to any mind-reading facilities.

[3] Such cases of total blocking are attested in the literature. For example, forms such as **furiosity*, **fallacity* do not exist because others (*fury*, *fallacy*) do. For more examples and discussion, see Blutner (1998).

[4] In the original formulation given in section 5.2.1, the recursion is indicated by the modulo clause in the Q- and R-principle.

[5] Levinson’s M-principle should not be confused with the markedness constraint M introduced in Table 5.1.

[6] For a short characterization of similarity, the reader is referred to Jäger (2007a). For a detailed treatment, see Tversky (1977).

[7] For details, the reader is referred to Winterstein (2011) who proposes an augmented argumentative approach for the contrast use.