

ZASPiL Nr. 51 – September 2009

Papers on Pragmasemantics

Anton Benz & Reinhard Blutner (Eds.)

Preface

Optimality theory as used in linguistics (Prince & Smolensky, 1993/2004; Smolensky & Legendre, 2006) and cognitive psychology (Gigerenzer & Selten, 2001) is a theoretical framework that aims to integrate constraint based knowledge representation systems, generative grammar, cognitive skills, and aspects of neural network processing. In the last years considerable progress was made to overcome the artificial separation between the disciplines of linguistic on the one hand which are mainly concerned with the description of natural language competences and the psychological disciplines on the other hand which are interested in real language performance.

The semantics and pragmatics of natural language is a research topic that is asking for an integration of philosophical, linguistic, psycholinguistic aspects, including its neural underpinning. Especially recent work on experimental pragmatics (e.g. Noveck & Sperber, 2005; Garrett & Harnish, 2007) has shown that real progress in the area of pragmatics isn't possible without using data from all available domains including data from language acquisition and actual language generation and comprehension performance. It is a conceivable research programme to use the optimality theoretic framework in order to realize the integration.

Game theoretic pragmatics is a relatively young development in pragmatics. The idea to view communication as a strategic interaction between speaker and hearer is not new. It is already present in Grice' (1975) classical paper on conversational implicatures. What game theory offers is a mathematical framework in which strategic interaction can be precisely described. It is a leading paradigm in economics as witnessed by a series of Nobel prizes in the field. It is also of growing importance to other disciplines of the social sciences. In linguistics, its main applications have been so far pragmatics and theoretical typology. For pragmatics, game theory promises a firm foundation, and a rigor which hopefully will allow studying pragmatic phenomena with the same precision as that achieved in formal semantics.

The development of game theoretic pragmatics is closely connected to the development of bidirectional optimality theory (Blutner, 2000). It can be easily seen that the game theoretic notion of a Nash equilibrium and the optimality theoretic notion of a *strongly* optimal form-meaning pair are closely related to each other. The main impulse that bidirectional optimality theory gave to research on game theoretic pragmatics stemmed from serious empirical problems that resulted from interpreting the principle of *weak* optimality as a synchronic interpretation principle.

In this volume, we have collected papers that are concerned with several aspects of game and optimality theoretic approaches to pragmatics.

The first paper about *Optimality-Theoretic Pragmatics* (Blutner and Zeevat) gives an overview about the application of OT to the domain of pragmatics. It reviews the three basic views – Relevance theory, Levinson’s theory of presumptive meanings, and the Neo-Gricean approach –, and it gives an optimality-theoretic restructuring of their core ideas. Further, it illustrates how bidirectional OT accounts for the synchronic and the diachronic perspective on pragmatic interpretation.

The second paper, *Optimality-Theoretic Pragmatics Meets Experimental Pragmatics* (Blutner), is discussing recent findings concerning the psychological reality of optimality-theoretic pragmatics. Further, the paper seeks to close the gap between experimental pragmatics and neo-Gricean theories of pragmatics.

Based on research by Smolensky and Gärdenfors, the paper entitled *Neural Networks, Penalty Logic and Optimality Theory* (Reinhard Blutner) is discussing the potential of OT as a theory that overcomes the gap between symbolic and neuronal systems. In the light of the proposed logical analysis notions like recoverability and bidirection are explained, and likewise the problem of founding a strict constraint hierarchy is discussed. Moreover, a claim is made for developing an “embodied” OT closing the gap between symbolic representation and embodied cognition.

The role of evolutionary strategies and signalling games for the development of natural language constructions is discussed in a joint paper by Tom Lentz and Reinhard Blutner (*Signalling Games and Optimal Constructions*). This paper is a reworking of an earlier squib written in Dutch.

Michael Franke’s contribution (*An Epistemic Interpretation of Bidirectional Optimality Based on Signalling Games*) is concerned with an epistemic interpretation of bidirectional optimality in terms of beliefs and strategies of players in a signalling game. In particular, the author demonstrates that strong optimality can be linked to an unsophisticated belief formation. Weak optimality, on the other hand is shown to correspond to higher-order iterated best response reasoning with an even more severe limitation on the belief formation process of agents.

In their paper *History and Grammaticalization of “doch”/“toch”*, Henk Zeevat and Elena Karagjosova compare the Dutch particle “toch” with the German pendant “doch”. As first noted by Doherty (1985), this comparison leads to a paradoxical question: If the sentence is presented with stress on “toch”/“doch” the conditions of use become the opposite from the same sentences with the stress removed. Zeevat and Karagjosova provide a new, historically based explanation of the paradox.

The paper by Anton Benz (*Outline of the Foundations for a Theory of Implicatures*) is an investigation into the foundations of the optimal answer approach as developed in (Benz, 2006; Benz & v. Rooij, 2007). It interprets the speaker’s signalling and the hearer’s interpretation behaviour as an objective natural regu-

larity. As natural regularity, communication can be described by causal Bayesian networks (Pearle, 2000). Benz uses this representation for explicating the notion of *common natural information*. From this notion, a general definition of implicature is derived. In the second part of the paper, this framework is extended to communication with efficient clarification requests and noisy speaker strategies.

References

- Benz, A. (2006). Utility and Relevance of Answers. In: A. Benz, G. Jäger und R. v. Rooij (eds.). *Game Theory and Pragmatics*. Basingstoke, Palgrave Mcmillan. S. 195–214.
- Benz, A. and van Rooij, R. (2007). Optimal Assertions and what they Implicate. *Topoi - an International Review of Philosophy*, 27(1), S. 63–78.
- Blutner, Reinhard (2000), Some aspects of optimality in natural language interpretation, *Journal of Semantics* 17, 189-216.
- Doherty, M. (1985). *Epistemische Bedeutung*. Berlin: Akademie-Verlag.
- Garrett, M., & Harnish, R. M. (2007). Experimental pragmatics: Testing for implicatures. *Pragmatics and Cognition*, 15, 65-90.
- Gigerenzer, G., & Selten, R. (2001). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, Mass: MIT Press.
- Grice, H.P. (1975). Logic and Conversation. In: P. Cole und J.L. Morgan (Hrsg.). *Syntax and semantics: Speech acts*. Vol 3, New York: Academic Press, S. 41–58.
- Noveck, I. A., & Sperber, D. (Eds.). (2005). *Experimental Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave MacMillan.
- Pearle, J. (2000). *Causality Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.
- Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Rutgers University and University of Colorado at Boulder: Technical Report RuCCSTR-2, available as ROA 537-0802. Revised version published by Blackwell, 2004.
- Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.

Berlin, November 2009

Anton Benz & Reinhard Blutner

Table of contents

<i>Reinhard Blutner & Henk Zeevat (University of Amsterdam)</i> Optimality-Theoretic Pragmatics	1
<i>Reinhard Blutner (University of Amsterdam)</i> Optimality-Theoretic Pragmatics Meets Experimental Pragmatics	27
<i>Reinhard Blutner (University of Amsterdam)</i> Neural Networks, Penalty Logic and Optimality Theory	53
<i>Tom Lentz (University of Utrecht) & Reinhard Blutner (University of Amsterdam)</i> Signalling Games: Evolutionary Convergence on Optimality	95
<i>Michael Franke (University of Amsterdam)</i> An Epistemic Interpretation of Bidirectional Optimality Based on Signalling Games	111
<i>Henk Zeevat (University of Amsterdam) & Elena Karagjosova (University of Stuttgart)</i> History and Grammaticalization of “doch” / “toch”	135
<i>Anton Benz (ZAS Berlin)</i> Outline of the Foundations for a Theory of Implicatures	153

Addresses of contributors

Anton Benz

Zentrum für Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18
10117 Berlin
Germany
Email: benz@zas.gwz-berlin.de

Reinhard Blutner

Institute for Language, Logic, and Computation (ILLC)
University of Amsterdam
P.O.Box 94242
1090 GE Amsterdam
The Netherlands
Email: blutner@uva.nl

Michael Franke

Institute for Language, Logic, and Computation (ILLC)
University of Amsterdam
P.O.Box 94242
1090 GE Amsterdam
The Netherlands
Email: m.franke@uva.nl

Elena Karagjosova

Institute for Linguistics
University of Stuttgart
P.O.Box 106037
70049 Stuttgart
Germany
Email: elena.karagjosova@ling.uni-stuttgart.de

Tom Lentz

Utrecht Institute of Linguistics (OTS)
University of Utrecht
Janskerhof 13
3512 BL Utrecht
The Netherlands
Email: T.O.Lentz@uu.nl

Henk Zeevat

Institute for Language, Logic, and Computation (ILLC)

University of Amsterdam

P.O.Box 94242

1090 GE Amsterdam

The Netherlands

Email: H.W.Zeevat@uva.nl

Optimality-Theoretic Pragmatics

Reinhard Blutner

ILLC, University of Amsterdam

Henk Zeevat

ILLC, University of Amsterdam

The article aims to give an overview about the application of Optimality Theory (OT) to the domain of pragmatics. In the introductory part we discuss different ways to view the division of labor between semantics and pragmatics. Rejecting the doctrine of literal meaning we conform to (i) semantic underdetermination and (ii) contextualism (the idea that the mechanism of pragmatic interpretation is crucial both for determining what the speaker says and what he means). Taking the assumptions (i) and (ii) as essential requisites for a natural theory of pragmatic interpretation, section 2 introduces the three main views conforming to these assumptions: Relevance theory, Levinson's theory of *presumptive meanings*, and the Neo-Gricean approach. In section 3 we explain the general paradigm of OT and the idea of bidirectional optimization. We show how the idea of optimal interpretation can be used to restructure the core ideas of these three different approaches. Further, we argue that bidirectional OT has the potential to account both for the synchronic and the diachronic perspective on pragmatic interpretation. Section 4 lists relevant examples of using the framework of bidirectional optimization in the domain of pragmatics. Section 5 provides some general conclusions. Modeling both for the synchronic and the diachronic perspective on pragmatics opens the way for a deeper understanding of the idea of naturalization and (cultural) embodiment in the context of natural language interpretation.

1 Introduction

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 & Legendre, 2006). In the study of natural language, OT was successfully applied to the main linguistic disciplines 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 & 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) is not without problems. This has to do with the normative character that is attributed to the Gricean setting. Speakers, as Grice puts 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. (Grice, 1975: 45)

It's obvious that this principle of cooperation is normative, and so are Grice's conversational maxims. If a person acts in a particular situation in a particular way we can ask *why* she did it the way she did; alternately, we can ask if it was *reasonable* what the person did, and if other options were possibly more reasonable in the given situation. Good Griceans are expected to ask the second type of questions whereas the first question is expected to be asked by cognitive scientists. While the normative and the naturalistic aspects of understanding human actions can be clearly separated from each other that does not mean they predict different action patterns in most cases. The idea of a rational world isn't so irrational to be excluded in 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 OT 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 the speaker better be cooperative or pretend to be cooperative if she wants to use language to bring about effects in hearers.

The naturalistic stance taken by OT pragmatics is one characteristic that brings it close to Relevance Theory (Sperber & Wilson, 1986/1995; Sperber & Wilson, 1995). Another point of agreement has to do with the way OT pragmatics views the division of labor between semantics and pragmatics. Taking the lead from Atlas (e.g. Atlas, 2005), both relevance theory (RT) and OT pragmatics reject the doctrine of literal meaning. And both approaches conform to the ideas of

- (i) semantic underdetermination

- (ii) contextualism (the suggestion that the mechanism of pragmatic interpretation is crucial both for determining what the speaker says and what he means).

In the broad view of OT, this framework can be seen as a general scheme that can be used for expressing many different and possibly diverging views. For instance, it is possible to give optimality-theoretic reconstructions of a speaker-oriented normative pragmatics like the one developed by Grice. It is also possible to reconstruct hearer-oriented naturalistic pragmatics as in RT (Hendriks & de Hoop, 2001; Zeevat, 2007b). These systems are important for online synchronic accounts of speaking and interpretation. But – perhaps most surprisingly – it is also possible to reconstruct the Neo-Gricean systems of Horn (1984), Atlas & Levinson (1981) and Levinson (2000). In contrast to RT where there is only one fundamental pragmatic principle (the presumption of optimal relevance), the Neo-Gricean systems have two opposing optimization principles, the Q- and the I-principle (Atlas and Levinson 1981, Horn 1984 who writes R instead of I) by two simultaneous optimization directions (the speaker and the hearer direction) and so obtain a bidirectional OT pragmatics. OT pragmatics in the narrower sense will start from this system and will show that Levinson's M-principle (iconicity) can be reduced to it. The system can also explain the emergence of mono-directional pragmatic systems that can account for online incremental interpretation in the style of RT. Given the divergences within the Neo-Gricean camp¹, it cannot be expected that a coherent theory like bidirectional OT-pragmatics can reconstruct all the views of all representatives of this camp.

The present chapter aims to give an overview of the application of OT to the domain of pragmatics. The assumptions (i) and (ii) are essential requisites for a natural theory of pragmatic interpretation. In section 2 we will introduce the three main views conforming to these assumptions: (a) RT, (b) Levinson's (2000) theory of *presumptive meanings*, and (c) the Neo-Gricean approach. In section 3 we explain the general paradigm of OT and the idea of bidirectional optimization. We show how the idea of optimal interpretation can be used to restructure the core ideas of these three different approaches. Further, we argue that bidirectional OT has the potential to account both for the synchronic and the diachronic perspective of pragmatic interpretation. Section 4 lists relevant examples of using the framework of bidirectional optimization in the domain of pragmatics. Section 5 provides some general conclusions. It argues that OT pragmatics has the potential to account both for the synchronic and the

¹ For instance, Horn (2005) points out that Levinson's (2000) view is very close in important respects to that of RT.

diachronic perspective in pragmatics. This bolsters the way for a deeper understanding of the idea of naturalization and (cultural) embodiment in the context of natural language interpretation.

2 The naturalization of pragmatics: three variations on Grice

The *naturalization of pragmatics* refers to a research program that aims to provide a cognitively realistic picture of utterance interpretation and production. Hence, the proponents of this program such as relevance theorists take the stance of seeing natural language interpretation as a cognitive phenomenon and thus considering the basic principles of communication as a consequence of the nature of human cognition. A prerequisite of this program deals with the levels of cognitive representations and the boundary between semantics and pragmatics. There is a strong tendency among current researchers to follow the tradition of radical pragmatics and to accept the following three claims:

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 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 LF than others). Obviously, this view sharply contrasts with the paradigm of Generative Semantics – a view that tries to ground pragmatic phenomena by using particular syntactic stipulations.

Before we come to a discussion of three variations on Grice and the naturalization of conversational implicatures in utterance interpretation it is useful first 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.’ (Chierchia 2004: 42). 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 Atlas & Levinson (1981), Gazdar (1979), (Horn, 1984), Soames (1982), Krifka (1995), Blutner (1998), Sauerland (2004), and Sæbø (2004); the local view is taken by Chierchia (2004), Levinson (2000), and Relevance Theory (Sperber & Wilson, 1986/1995; e.g. Carston, 2002).

Usually, the globalists argue against the local view and the localists against the global view. We 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, we will argue that seemingly conflicting approaches such as relevance theory and the neo-Gricean approach are much closer related than expected by its opponents. In section 3, once more OT will prove his power of unification in giving hints 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 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:

- (i) to be at least relevant enough to warrant the addressee’s processing effort
- (ii) 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, & Girotto, 1995: 95): (a) test possible interpretations in their order of accessibility, (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:

- (i) Differing from both RT and the standard neo-Gricean view, Levinson assumes *three* levels of meaning corresponding to sentence meaning, utterance-type meaning and utterance-token meaning
- (ii) utterance-type meanings are in correspondence with Grice' 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*
- (iii) In contrast to Grice' 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-heuristics). The feature of local pragmatics is essential to artificial intelligence pragmatics (e.g. Hobbs & 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 & Levinson, 1981; Horn, 1984; Blutner, 1998; e.g. Atlas, 2005; Horn, 2005) are assuming two countervailing optimization principles: the Q-principle and the R-principle.² 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, 2005):

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)

² 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.

(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-heuristics. 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.³ The difference will become quite clear in the following section when we give formalization in terms of bidirectional OT.

Sometimes it is stressed that there is a fundamental difference in perspective and goals between the neo-Gricean and the RT approaches to pragmatics. For instance, Horn (2005) claims the following:

Grice's goal of developing an account of speaker meaning (of which implicature constitutes a proper subpart) is distinct from Relevance theorists' goal of developing a cognitive psychological model of utterance interpretation, which does not address the question of how and why the speaker, given what she wants to convey, utters what she utters. (194).

This seems to express the difference between the naturalistic stance and the normative stance mentioned in section 1. However, we agree with Carston (2005) that this statement is too strong as it stands since RT (as does Horn's theory) makes some predictions about **why** the speaker, given her communicative intention, utters what she utters. Further, the difference between the normative stance and the naturalistic stance should not be overestimated because in practice there is seldom a deep empirical conflict between the two

³ 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.

stances (Spohn, 1993). A much more important question concerns the status of the theory with regard to synchrony versus diachrony. Both RT and Levinson's theory of *presumptive meaning* takes the synchronic view where neo-Griceans take both views (and, sometimes, confuse them). In the following section we will see how OT relates both views/perspectives.

3 The framework of OT

OT can be seen as a general framework that systematizes the use of optimization methods in linguistics.⁴ One component of OT is a list of tendencies that hold for observable properties of a language. These tendencies take the form of violable constraints. Because the constraints usually express very general statements, they can be in conflict. Conflicts among constraints are resolved because the constraints differ in strength. Minimal violations of the constraints (taking their strength into account) define optimal conflict resolutions.

Standardly, OT specifies the relation between an input and an output. This relation is mediated by two formal mechanisms, **GEN** and **EVAL**. GEN (for Generator) creates possible output candidates on the basis of a given input. EVAL (for Evaluator) uses the particular constraint ranking of the universal set of constraints **CON** to select the best candidate for a given input from among the candidate set produced by GEN. In phonology and syntax, the input to this process of optimization is an underlying linguistic representation. The output is the (surface) form as it is expressed. Hence, what is normally used in phonology and syntax is unidirectional optimization. Obviously, the point of view of the speaker is taken. This contrasts with OT semantics where the view of the hearer is taken (de Hoop & de Swart, 2000; Hendriks & de Hoop, 2001).

Bidirectional optimization (Blutner, 1998, 2000) integrates the speaker and the hearer perspective into a simultaneous optimization procedure. In pragmatics, this bidirectional view is motivated by a reduction of Grice's maxims of conversation to two principles: the R-principle, which can be seen as the force of unification minimizing the Speaker's effort, and the Q-principle, which can be seen as the force of diversification minimizing the Auditor's effort. In a slightly different formulation, the R-principle seeks to select the most coherent interpretation⁵ and the Q-principle acts as a blocking mechanism which blocks all the outputs which can be expressed more economically by an alternative linguistic input. This formulation makes it quite clear that the neo-

⁴ A recent overview is given in Smolensky & Legendre (2006). For OT pragmatics the reader is referred to Blutner & Zeevat (2004) and Blutner, de Hoop & Hendriks (2005).

⁵ What is meant by coherence has to be expressed by particular OT constraints, such as formulated by Zeevat (2007a, 2007b) for instance.

Gricean framework can be conceived of as a bidirectional optimality framework which integrates the speaker and the hearer perspective. Whereas the R-principle compares different possible interpretations for the same syntactic expression, the Q-principle compares different possible syntactic expressions that the speaker could have used to communicate the same meaning.

We will give a very schematic example in order to illustrate some characteristics of the bidirectional OT. Assume that we have two forms f_1 and f_2 which are semantically equivalent. This means that **GEN** associates the same interpretations with them, say m_1 and m_2 . We stipulate that the form f_1 is less complex (less marked) than the form f_2 and that the interpretation m_1 is less complex (less marked) than the interpretation m_2 . This is expressed by two markedness constraints: F for forms and M for interpretations – F prefers f_1 over f_2 and M prefers m_1 over m_2 . This is indicated by the two leftmost constraints in table (1).

Table 1: Markedness and bias constraints in a 2-forms \times 2-interpretations design

	F	M	F→M	*F→*M	F→*M	F*→M
$\langle f_1, m_1 \rangle$					*	
$\langle f_1, m_2 \rangle$		*	*			
$\langle f_2, m_1 \rangle$	*			*		
$\langle f_2, m_2 \rangle$	*	*				*

Besides the markedness constraints, four so-called linking constraints can be formulated. There are precisely four independent linking constraints in the present example. The linking constraint F→M says that simple (unmarked) forms express simple interpretations. Hence, this is a straightforward formalization of Levinson’s (2000) **I**-heuristics as an OT constraint. The constraint *F→*M says that complex forms express complex interpretations, and this is an expression of Levinson’s **M**-heuristics⁶. The two remaining bias 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. With only two forms and two meanings, the

⁶ Levinson’s M-principle should not be confused with the markedness constraint M introduced in Table 1.

substance of the Q-heuristics is not really different from that of the M-constraint.⁷

In the so-called strong version of bidirectional OT, a form-interpretation pair $\langle f, m \rangle$ is considered to be (strongly) optimal iff

- Interpretive Optimization: no other pair $\langle f, m' \rangle$ can be generated that satisfies the constraints better than $\langle f, m \rangle$ and
- Expressive Optimization: no other pair $\langle f', m \rangle$ can be generated that satisfies the constraints better than $\langle f, m \rangle$.

From the differences of markedness given by the constraints F and M the ordering relation between form-meaning pairs can be derived as shown in Figure 1. The preferences are indicated by arrows in a two-dimensional diagram. Such diagrams give an intuitive visualization for the optimal pairs of (strong) bidirectional OT: they are simply the meeting points of horizontal and vertical arrows.⁸ The optimal pairs are marked with the symbol ✂ in the diagram.

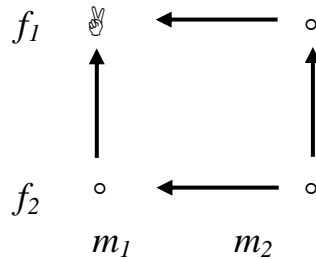


Figure 1: Diagram to illustrate strong bidirection

The scenario just mentioned describes the case of *total blocking* where some forms (e.g., *furiousness, *fallaciousness) do not exist because others do (fury, fallacy). However, blocking is not always total but may be partial. This means that not all the interpretations of a form must be blocked if another form exists. McCawley (1978) collects a number of examples demonstrating the phenomenon of *partial blocking*. For example, he observes that the distribution of productive causatives (in English, Japanese, German, and other languages) is

⁷ Harmonic alignment (Prince & Smolensky 1993, Aissen 2003) is precisely the fact that these two linking constraints hold.

⁸ Dekker & van Rooy (2000), who introduced these diagrams, gave bidirectional OT a game theoretic interpretation where the optimal pairs can be characterized as so-called Nash Equilibria.

restricted by the existence of a corresponding lexical causative. Whereas lexical causatives (e.g. (1a)) tend to be restricted in their distribution to the stereotypical causative situation (direct, unmediated causation through physical action), productive (periphrastic) causatives 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

To make things concrete we can take f_1 to be the lexical causative form (1a), f_2 the periphrastic form (1b), m_1 direct (stereotypic) causation and m_2 indirect causation.

Typical cases of partial blocking are found in morphology, syntax and semantics. The general tendency of partial blocking seems to be that "unmarked forms tend to be used for unmarked situations and marked forms for marked situations" (Horn 1984: 26) – a tendency that Horn (1984: 22) terms "*the division of pragmatic labour*".

There are two ways of avoiding total blocking within the bidirectional OT framework and to describe Horn's division of pragmatic labour. The first possibility makes use of linking constraints and fits the intended form-interpretation relation by stipulating the appropriate ranking of the constraints such that partial blocking comes out. Let's assume that the two bias-constraints $F \rightarrow M$ and $*F \rightarrow *M$ are higher ranked than the rest of the constraints. This can be depicted as in Figure 2a. Hence, strong bidirection can be taken as describing Horn's division of pragmatic labour when the appropriate linking constraints are dominating.

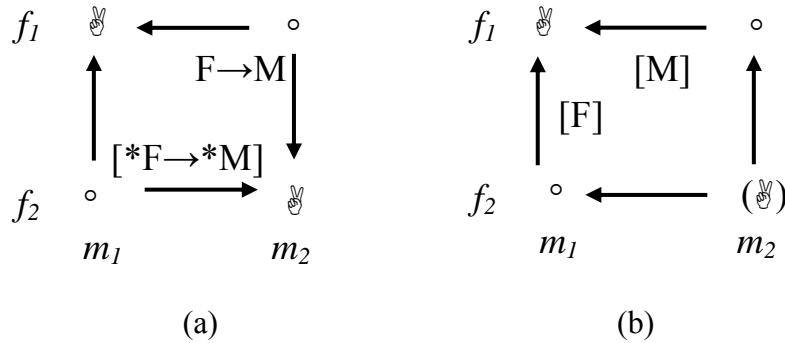


Figure 2: Two ways of describing Horn's division of pragmatic labour: (a) by assuming two dominant bias constraints; (b) by assuming markedness constraints and weak bidirection

The second possibility is to weaken the notion of (strong) optimality in a way that allows us to derive Horn's division of pragmatic labour by means of the *evaluation procedure* and without stipulating particular bias constraints. Blutner (2000) proposes a *weak* version of two-dimensional OT, according to which the two dimensions of optimization are mutually related:

Super-Optimality

A form-interpretation pair $\langle f, m \rangle$ is called super-optimal iff

- Interpretive Optimization: no other super-optimal pair $\langle f, m' \rangle$ can be generated that satisfies the constraints better than $\langle f, m \rangle$;
- Expressive Optimization: no other super-optimal pair $\langle f', m \rangle$ can be generated that satisfies the constraints better than $\langle f, m \rangle$.

This formulation looks like a circular definition, but Jäger (2002) has shown that this is a sound *recursive* definition under very general conditions (well-foundedness of the ordering relation). The important difference between the weak and strong notions of optimality is that the weak one accepts super-optimal form-meaning pairs that would not be optimal according to the strong version. It typically allows marked expressions to have an optimal interpretation, although both the expression and the situations they describe have a more efficient counterpart.

Figure 2b shows that the *weak* version of bidirection can explain the effects of partial blocking without the stipulation of extra bias constraints; especially it can explain why the marked form f_2 gets the marked interpretation m_2 . This is a

consequence of the *recursion* implemented in weak bidirection:⁹ the pairs $\langle f_1, m_2 \rangle$ and $\langle f_2, m_1 \rangle$ are *not* super-optimal. Hence, they cannot block the pair $\langle f_2, m_2 \rangle$ and it comes out as a new super-optimal pair. In this way, the weak version accounts for Horn's pattern of *the division of pragmatic labour*.

The two parts of Figure 2 describe the same set of solution pairs but the calculation of the solutions is completely different in the two cases. In the first case unidirectional optimization (either hearer or speaker perspective) is sufficient to calculate the solution pairs. It is plausible to assume that this kind of OT systems can be used to construct cognitively realistic models of online, incremental interpretation (cf. Blutner 2006). The second case – using the recursion of weak bidirection (super-optimality) – has a completely different status. Because of its strictly non-local nature the proposed algorithm that calculate the super-optimal solutions do not even fit the simplest requirements of a psychologically realistic model of online, incremental interpretation (Zeevat, 2000; Beaver & Lee, 2004)¹⁰. The proper understanding of weak bidirection relates best to an off-line mechanism that is based on bidirectional learning (Blutner, Borra, Lentz, Uijlings, & Zevenhuijzen, 2002; Benz, 2003b; Van Rooy, 2004). In these approaches the solution concept of weak bidirection is considered as a principle describing the results of language change: super-optimal pairs emerge over time in language change. This relates to the view of Horn (1984) who considers the Q and the I principle as diametrically opposed forces in language change. This conforms to the good old idea that synchronic structure is significantly informed by diachronic forces.

For the sake of illustration let's go back to our example illustrated in (1). Let's assume a population of agents who realize speaker- and hearer strategies based exclusively on the markedness constraints F and M. In this population each content is expressed in the simplest way (f_1) and each expression is understood in the simplest way (m_1). Let's assume further that these agents communicate with each other. When agent x is in the speaker role and intends to express m_1 , then expressive optimization yields f_1 . Agent y is a hearer who receives f_1 and, according to interpretive optimization, he gets the interpretation m_1 – hence the hearer understands what the speaker intends: successful communication. Now assume the speaker wants to express m_2 . With the same

⁹ In the original formulation given in section 2, the recursion is indicated by the modulo-clause.

¹⁰ There are several arguments why bidirectional OT cannot yield an online mechanism of linguistic competence. Beaver & Lee (2004) argue that if more rounds of optimization are allowed, the bidirectional OT-model severely overgenerates in the sense that in later rounds peculiar new form-meaning pairs will emerge as winners. Before the Beaver & Lee paper, Zeevat (2000) argued against the symmetric view of OT pragmatics starting from the famous rat/rad problem and its pragmatic counterparts.

logic of optimization he will produce f_I and the agent y interprets it as m_I . In this case, obviously, the communication is not successful. Now assume some kind of *adaptation* either by iterated learning or by some mutations of the ranked constraint system (including the bias constraints). According to this adaptation mechanism the expected ‘utility’ (how well they understand each other in the statistical mean) can improve in time. In that way a system that is evolving in time can be described including its special attractor dynamics. In each case there is a stabilizing final state that corresponds to the system of Figure 2a where the two Levinsonian (2000) constraints **I** ($= [F \rightarrow M]$) and **M** ($= [F \rightarrow M]$) outrank the rest of the constraints. It is precisely this system that reflects Horn’s division of pragmatic labour. The only condition we have to assume is that the marked contents are less frequently expressed than the unmarked contents.¹¹

Hence, the important insight is that a system that is exclusively based on markedness constraints such as in Figure 2b is evolutionary related to a system based on highly ranked bias constraints such as in Figure 2a. We will use the term *fossilization* for describing the relevant transfer.¹²

Now we come back to the earlier goal of giving an OT reconstruction of the three variations on Grice (section 2). 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 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 to this theory and can be used for a straightforward formalization. Unsurprisingly, this conception can be seen best from a diachronic perspective, at least so far 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 never was 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’s 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

¹¹ For more discussion of the role of frequencies in an evolutionary setting cf. Stalnaker (2006). The general conclusion is that the solution concept of weak bidirection can be seen as a rough first approximation to the more adequate solution concepts of evolutionary game theory that describe the results of language change.

¹² In a somewhat different context, Peter Cole (Cole, 1975) calls it “lexicalization of contextual meaning”.

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, 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 4) in the light of the general structure of RT – a task that goes beyond what can be done in the present paper.¹³

We have mentioned already that there is a relation between diachronic and synchronic systems, and we have introduced the term *fossilization* for describing the relevant transfer.¹⁴ Taken the existence of this transfer, it can be demonstrated that the three discussed variations on Grice are much closer related than the occasional polemics let us expect.

In order to get an impression of OT pragmatics at work we shortly will discuss Zeevat's (2002, 2007a, 2007b) reconstruction of presupposition theory as formulated by Van der Sandt (1992) and Heim (1983). In both these theories there are two defaults or preferences. The first one prefers identifying the induced presupposition in the context of the utterance (resolution), the second one prefers the addition of the presupposition to the global context (Heim) or to the highest accessible context where that is possible (Van der Sandt). The reconstruction makes use of the following constraints that are used in finding an optimal interpretation: FAITH > CONSISTENCE > DO NOT ACCOMMODATE > STRENGTH. CONSISTENCE demands that there is no conflict of the current utterance with what is known already, FAITH asks for the presence of the presupposed information at an accessible position, DO NOT ACCOMMODATE forbids the addition of the presupposed information and STRENGTH forbids interpretations if there are informationally stronger ones available. The OT system improves in several ways on the theories that it tries to reconstruct. DO NOT ACCOMMODATE prefers partial resolutions to full accommodations and does not militate against copies of presuppositions. STRENGTH often gives a better

¹³ Another important aspect concerns pragmatic acceptability. The RT account of pragmatic (un)acceptability is carefully worked out in connection with bridging phenomena (Wilson & Matsui, 1998). In RT, "unacceptability can result from (a) inadequate effects or (b) gratuitous effort" ((Wilson & Matsui, 1998: 19). That means there have to be thresholds for (a) effects and (b) effort, and when these thresholds are reached unacceptability results. This kind of argumentation is not possible within an OT approach because in OT the absolute strength of constraint violation is not of importance. What counts is the comparison with other expressions that lead to the same interpretation and the possibility of blocking an interpretation by a cheaper expression alternatives (e.g. Blutner, 1998). It's an open issue what are the empirical consequences of this view in case of bridging.

¹⁴ Concerning the debate whether certain pragmatic inferences are truly conversational or whether they have become lexically encoded the reader is referred to Cole (1975) and Potts (to appear).

prediction of the accommodation site than van der Sandt (1992). Zeevat (2002) uses the reconstruction to explore particles like "too". These particles have an exceptional behaviour within accounts of presupposition: they do not allow (full) accommodation and are obligatory in the contexts in which they occur. The second phenomenon needs a max(F) constraint as in OT phonology: certain relations to the context need to be marked. But the other phenomenon seems to allow a bidirectional explanation. DO NOT ACCOMMODATE in a bidirectional interpretation prohibits candidate expressions that force accommodations, if there is a simple alternative that means the same and does not force the accommodation. For particles, the sentence without the particle always is an alternative of this kind.

4 Some applications of bidirectional OT in pragmatics

OT pragmatics has been used for describing a series of phenomena and observation in the domain of natural language interpretation. This section gives an overview of some of these applications without going into any technical or empirical details.

- *Centering theory.* (Beaver, 2004) is using bidirectional OT as framework for the reformulation of centering models of pronoun resolution.
- *Disambiguation.* Gärtner (2004b, 2004a) analyses Icelandic object-shift and differential marking of (in-)definites in Tagalog addressing the issue of disambiguation and partial iconicity in natural language.
- *Differential object marking:* Aissen (2003), Nilsenova (2002), and Jäger & Zeevat (2002) discuss the relevant correlation between grammatical functions and semantic/pragmatic properties.
- *Binding theory.* Mattausch (2004a, 2004b) introduces the influential work of Levinson on the origin and typology of binding theory (summarized in Levinson, 2000) 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. For early work on discourse anaphora in a bidirectional framework we refer to Buchwald, Schwartz, Seidl & Smolensky (2002).
- *Pragmatics for propositional attitudes.* Aloni (2001, 2005b) has argued that a number of seeming paradoxes emerging from logical analyses of attitudes and questions can be explained in terms of shifts in perspective over the universe of discourse. Shifts in perspective have a cost and, therefore, are generally avoided. However, under certain circumstances such shifts are

required to comply with general principles of rational conversation, which, for example, disallow vacuous or inconsistent interpretations. Aloni's work suggests a formulation of a perspective selection procedure in the framework of bidirectional OT.

- *Discourse particles and presupposition.* Zeevat (2002, 2004) treats discourse particles within an extended OT reconstruction of presupposition theory and concludes that more particles can be treated and the analysis becomes simpler if one starts from the fact that discourse particles are obligatory if the context of utterance and the current utterance stand in one of a number of special relations, like adversativity, additivity, contrast, etc. In another paper, Zeevat (2007a) provides a full solution to the projection problem for presuppositions. Jäger & Blutner (2000, 2003) suggest an bidirectional analysis of the different reading of German 'wieder' (again).
- *Complex implicatures.* Blutner (2007) gives an OT account of implicature projection and explains the relevance 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, & Hindsill, 2007)
- *Marking and Interpretation of negation.* Henriëtte de Swart (2004) provides a bi-directional OT approach to the syntax and pragmatics of negation and negative indefinites (see also, de Swart, in press).
- *Scalar implicatures and exhaustification.* Exhaustivity implicatures and also scalar implicatures depend on the issue under discussion which can be formalized using Groenendijk & Stokhof's theory of question and answers. Combining this framework with those of bidirectional OT, Aloni (2005b, 2005a) explains several puzzles in this area.
- *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).
- *Stage level/individual level contrast.* Maienborn (2004, 2005) argues against the popular view that the distinction between *stage level predicates* and *individual level predicates* rests on a fundamental cognitive division of the world that is reflected in the grammar. Instead, Maienborn proposes a pragmatic explanation of the distinction, and she gives, inter alia, a discourse-based account of Spanish *ser/estar*.
- *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. Well-studied examples include narrowing (e.g. drink used to mean 'alcoholic drink'), approximation (e.g. square used to mean 'squarish') and metaphorical extension (e.g. battleaxe used to mean 'frightening person'). Lexical Pragmatics can be formulated by using the formal instruments of OT-based pragmatics (Blutner, 1998; Blutner et al., 2005). Prototypical applications include the pragmatics of dimensional adjectives (Blutner & Solstad, 2000), the analysis of Dutch om/rond (Zwarts, 2006), the pragmatics of negated antonyms (Blutner, 2004; Krifka, to appear), the approximate interpretation of number words (Krifka, 2007), several examples of semantic change (Eckardt, 2002).
- *Language acquisition and learning*: There are several studies that test the role of weak bidirection in developing interpretation and production preferences in connection with indefinite NPs (deHoop & Kramer, 2005/2006) and pronominal anaphors (Hendriks & Spender, 2005/2006; Hendriks, Rijn, & Valkenier, 2007; Mattausch & Gülzow, 2007). From a theoretical perspective, the problem of learning is investigated by Benz (2003a, 2003b).

5 Conclusions

The error in many formulations of pragmatic inferences is that synchrony and diachrony are confused. OT pragmatics accounts both for the synchronic perspective – by formulating a localist, incremental model based on unidirectional optimization using a emerging system of linking constraints – and the diachronic perspective – using the solution concept of weak bidirection which conforms to a strictly global view. The perspectives are connected by the idea of fossilization.

Many patterns in language have been proposed to be directly or indirectly influenced by the conflict between multiple influences on output form. Within phonology for example, the notion that conflict between minimization of articulatory effort and maximization of perceptual distinctiveness has an influence on grammatical patterns has held currency at least since Baudouin de Courtenay (1895). Contemporary work grounding phonological patterns in optimization of conflicting influences on output form include work done within Natural Phonology (Stampe, 1973), Grounded Phonology (Archangeli & Pulleyblank, 1994), and Optimality Theory (Prince & Smolensky, 1993/2004; Boersma, 1998) to name but a few. Weak bidirection describes the interaction of these forces in an approximate but simply to understand way. However, for fully understanding the bidirectional game that leads to the resolution of the conflict

between the opposite forces, evolutionary game theory provides a more adequate model (e.g. Van Rooy, 2004).

Relevance Theory and Levinson's theory of presumptive meaning account for the resolution of the conflict between effort minimization and effect maximization in different ways. In a certain sense, the crux of both approaches can be translated in OT pragmatics by making use of particular linking constraints. This translation makes the advantage of both approaches visible: both conform to the incremental, online character of natural language interpretation.¹⁵

We have argued that OT pragmatics has the potential to account both for the synchronic and the diachronic perspective in pragmatics, and for the way both are related to each other. We further have pointed out that the concepts of fossilization can help to understand the idea of naturalization and (cultural) embodiment in the context of natural language interpretation. However, there are also important open questions regarding the status of fossilization. In a by now classical paper Cole (1975) considered the following example of a true conversational implicature, where a girl called Pamela upon being asked (2) might reply (3):

- (2) How are you doing in your new position at San Andreas Fault University?
(3) Well, I haven't been fired yet.

Although the logical content of (3) is roughly that of the proposition that Pamela has not yet lost her job, more than that is implicated, namely that Pamela is not doing well. In this example, the implicature is really novel. There is no construction involved whose frequent use could lead to the fossilization phenomenon (Cole's term is 'lexicalization'). Hence, this implicature is different from many other cases where a certain amount of fossilization is plausible. The important question is how to discriminate between offline implicatures that are not fossilized and their fossilized counterparts. Where is the boundary between aspects of interpretations that are truly conversational and aspects which have become lexically (or syntactically) encoded? We think the former aspect of interpretation can require some real mind reading capacities,

¹⁵ In discussing processing characteristics, incrementality and automaticity of processing have to be discriminated. Whereas automaticity of processing implicates the incremental character of processing the opposite is not true: incrementality does not implicate automatic processing. RT explains the incremental character of processing and has good reasons for assuming controlled processing in order to account for the processing of conversational implicatures. That's different from Levinson's account which assumes automatic processing for generalized conversational implicatures. It seems that RT is better justified on empirical grounds (cf. Noveck & Sperber, 2005).

requires conscious reflections and proceeds offline. So far we can see, none of the discussed pragmatic theories has an interesting answer for this long-standing and intriguing question.

6 References

- Aissen, J. (2003). Differential coding, partial blocking, and bidirectional OT. In P. Nowak & C. Yoquelet (Eds.), *Proceedings of the 29th Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society.
- Aloni, M. (2001). Pragmatics for Propositional Attitudes. In R. v. Rooy & M. Stokhof (Eds.), *Proceedings of the Thirteenth Amsterdam Colloquium*. Amsterdam: University of Amsterdam.
- Aloni, M. (2005a). *Expressing ignorance or indifference. Modal implicatures in BiOT* Paper presented at the Batumi 2005 symposium Batumi.
- Aloni, M. (2005b). A Formal Treatment of the Pragmatics of Questions and Attitudes. *Linguistics and Philosophy*, 28, 505-539.
- Aloni, M., Butler, A., & Hindsill, D. (2007). Nuclear Accent in Bidirectional Optimality Theory. In M. Aloni & A. Butler & P. Dekker (Eds.), *Questions in Dynamic Semantics*: Elsevier.
- Archangeli, D. B., & Pulleyblank, D. (1994). *Grounded Phonology*. Cambridge MA: MIT Press.
- Atlas, J. D. (2005). *Logic, Meaning, and Conversation: Semantical Underdeterminacy, Implicature, and Their Interface*. Oxford: Oxford University Press.
- Atlas, J. D., & Levinson, S. C. (1981). It-clefts, informativeness and logical form. In P. Cole (Ed.), *Radical Pragmatics* (pp. 1-61). New York: Academic Press.
- Axelrod, R. (1984). *The evolution of co-operation*. London: Penguin.
- Baudouin de Courtenay, J. (1895). *Versuch einer Theorie phonetischer Alternationen: Ein Capitel aus der Psychophonetik*. Strassburg/Crakow: Trubner.
- Beaver, D. (2004). The optimization of discourse anaphora. *Linguistics and Philosophy*, 27, 3-56.
- Beaver, D., & Lee, H. (2004). Input-output mismatches in OT. In Palgrave/Macmillan (Ed.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire.
- Benz, A. (2003a). *Partial blocking and associative learning*. Unpublished manuscript, Berlin & Kolding. Available from <http://www.anton-benz.de/Paper.html>.
- Benz, A. (2003b). Partial Blocking, associative learning, and the principle of weak optimality. In J. Spenader & A. Eriksson & Ö. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory* (pp. 150-159). Stockholm.
- Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, 15, 115-162.

- Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17, 189-216.
- Blutner, R. (2004). Pragmatics and the lexicon. In L. Horn & G. Ward (Eds.), *Handbook of pragmatics*. Oxford: Blackwell.
- Blutner, R. (2006). Embedded implicatures and optimality theoretic pragmatics. In Torgim Solstad & A. Grønn & D. Haug (Eds.), *A Festschrift for Kjell Johan Sæbø: in partial fulfilment of the requirements for the celebration of his 50th birthday*. Oslo.
- Blutner, R. (2007). Optimality Theoretic Pragmatics and the Explicature/Implicature Distinction. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 67-89). Houndmills, Basingstoke, Hampshire: Palgrave/MacMillan.
- Blutner, R., Borra, E., Lentz, T., Uijlings, A., & Zevenhuijzen, R. (2002). Signalling games: Hoe evolutie optimale strategieën selecteert, *Handelingen van de 24ste Nederlands-Vlaamse Filosofiedag*. Amsterdam: Universiteit van Amsterdam.
- Blutner, R., de Hoop, H., & Hendriks, P. (2005). *Optimal Communication*. Stanford: CSLI Publications.
- Blutner, R., & Solstad, T. (2000). Dimensional designation: a case study in lexical pragmatics. In R. Blutner & G. Jäger (Eds.), *Studies in Optimality Theory* (pp. 30-40). Potsdam: University of Potsdam.
- Blutner, R., & Zeevat, H. (Eds.). (2004). *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Boersma, P. (1998). *Functional phonology*. The Hague: Holland Academic Graphics.
- Buchwald, A., Schwartz, O., Seidl, A., & Smolensky, P. (2002). Optimality Theory: Discourse Anaphora in a Bidirectional framework citeseer.ist.psu.edu/buchwald02optimality.html.
- Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.
- Carston, R. (2005). Relevance Theory, Grice and the neo-Griceans: a response to Laurence Horn's 'Current issues in neo-Gricean pragmatics'. *Intercultural Pragmatics*, 2, 303-319.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond* (pp. 39-103). Oxford: Oxford University Press.
- Cole, P. (1975). The synchronic and diachronic status of conversational implicature. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Volume 3: Speech Acts* (pp. 257-288). San Diego, Cal.: Academic Press, Inc.
- de Hoop, H. (2004). On the interpretation of stressed pronouns. In R. Blutner & H. Zeevat (Eds.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- de Hoop, H., & de Swart, H. (2000). Temporal adjunct clauses in optimality theory. *Rivista di Linguistica*, 12, 107-127.

- de Swart, H. (2004). Marking and Interpretation of negation: A bi-directional OT approach. In R. Zanuttini & H. Campos & E. Herburger & P. Portner (Eds.), *Negation, Tense and Clausal Architecture: Cross-linguistic Investigations*: Georgetown University Press.
- de Swart, H. (in press). *Expression and interpretation of negation* (Vol. 77). Dordrecht: Springer.
- de Hoop, H., & Kramer, I. (2005/2006). Children's Optimal Interpretations of Indefinite Subjects and Objects. *Language Acquisition*, 13, 103-123.
- Dekker, P., & van Rooy, R. (2000). Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, 17, 217-242.
- Eckardt, R. (2002). Semantic Change in Grammaticalization. In G. Katz & S. Reinhard & P. Reuter (Eds.), *Proceedings of the Sixth Annual Meeting of the Gesellschaft für Semantik, Sinn & Bedeutung VI*. Osnabrück: University of Osnabrück.
- Fodor, J. (1975). *The Language of Thought*. New York: Thomas Crowell.
- Gärtner, H.-M. (2004a). On Object-Shift in Icelandic and Partial Iconicity. *Lingua*, 114, 1235-1252.
- Gärtner, H.-M. (2004b). On the OT-status of 'unambiguous encoding'. In R. Blutner & H. Zeevat (Eds.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Gazdar, G. (1979). *Pragmatics*. New York: Academic Press.
- Grice, P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics*, 3: *Speech Acts* (pp. 41-58). New York: Academic Press.
- Heim, I. (1983). File change semantics and the familiarity theory of definiteness, *Meaning, Use and the Interpretation of Language* (pp. 164-190). Berlin: Walter de Gruyter.
- Hendriks, P. (2004). Optimization in Focus Identification, *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Hendriks, P., & de Hoop, H. (2001). Optimality theoretic semantics. *Linguistics and Philosophy*, 24, 1-32.
- Hendriks, P., Rijn, H. v., & Valkenier, B. (2007). Learning to reason about speakers' alternatives in sentence comprehension: A computational account. *Lingua*, 117, 1879-1896.
- Hendriks, P., & Spenader, J. (2005/2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, 13, 319-348.
- Hobbs, J., & Martin, P. (1987). Local Pragmatics, *Proceedings, International Joint Conference on Artificial Intelligence* (pp. 520-523). Milan.
- Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.
- Horn, L. (1989). *A natural history of negation*. Chicago: Chicago University Press.

- Horn, L. (2004). Implicature. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics*. Oxford: Blackwell.
- Horn, L. (2005). Current issues in neo-Gricean pragmatics. *Intercultural Pragmatics*, 2, 191-204.
- Jäger, G. (2002). Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information*, 11, 427-451.
- Jäger, G., & Blutner, R. (2000). Against lexical decomposition in syntax. In A. Z. Wyner (Ed.), *Proceedings of the Fifteenth Annual Conference, IATL 7* (pp. 113-137). Haifa: University of Haifa.
- Jäger, G., & Blutner, R. (2003). Competition and interpretation: The German adverb wieder ("again"). In C. Fabricius-Hansen & E. Lang & C. Maienborn (Eds.), *Handbook of Adjuncts* (pp. 393-416). Berlin: Mouton de Gruyter.
- Jäger, G., & Zeevat, H. (2002). A reinterpretation of syntactic alignment. In D. d. Jongh & H. Zeevat & M. Nilsenova (Eds.), *Proceedings of the 3rd and 4th International Symposium on Language, Logic and Computation*. Amsterdam: ILLC.
- Krifka, M. (1995). The Semantics and Pragmatics of Polarity Items. *Linguistic Analysis*, 25, 209-257.
- Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma & I. Krämer & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 111-126). Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen.
- Krifka, M. (to appear). Negated antonyms: Creating and filling the gap. In U. Sauerland & P. Stateva (Eds.), *NN*.
- Levinson, S. (2000). *Presumptive meaning: The theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.
- Maienborn, C. (2004). A pragmatic explanation of the stage level/individual level contrast in combination with locatives. In B. Agbayani & V. Samiian & B. Tucker (Eds.), *Proceedings of the Western Conference on Linguistics (WECOL)*, volume 15 (pp. 158 - 170). Fresno: CSU.
- Maienborn, C. (2005). A discourse-based account of Spanish ser/estar. *Linguistics*, 43, 155-180.
- Mattausch, J. (2004a). *On the Optimization & Grammaticalization of Anaphora*. Unpublished Ph.D. Thesis, Humboldt University, Berlin.
- Mattausch, J. (2004b). Optimality Theoretic Pragmatics and Binding Phenomena. In R. Blutner & H. Zeevat (Eds.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Mattausch, J., & Gülzow, I. (2007). A note on acquisition in frequency-based accounts of Binding Phenomena. In I. Gülzow & N. Gagarina (Eds.), *Frequency Effects in Language Acquisition: Defining the Limits of Frequency as an Explanatory Concept* (pp. 331-357). Berlin. New York: Mouton de Gruyter.

- McCawley, J. D. (1978). Conversational implicature and the lexicon. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 245-259). New York: Academic Press.
- Nilsenova, M. (2002). The Pragmatics of Differential Object Marking. In M. Nissim (Ed.), *Proceedings of the ESSLLI '02 Student Session*. Trento, Italy: University of Trento, revised version available from <http://staff.science.uva.nl/~mnilseno/>.
- Noveck, I. A., & Sperber, D. (Eds.). (2005). *Experimental Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave MacMillan.
- Potts, C. (to appear). Into the conventional-implicature dimension. *Philosophy Compass*.
- Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Rutgers University and University of Colorado at Boulder: Technical Report RuCCSTR-2, available as ROA 537-0802. Revised version published by Blackwell, 2004.
- Sæbø, K. J. (2004). Optimal interpretations of permission sentences. In D. de Jongh & P. Dekker (Eds.), *Proceedings of the 5th Tbilisi Symposium on Language, Logic and Computation* (pp. 137-144). Amsterdam and Tbilisi.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27, 367-391.
- Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.
- Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry*, 13, 483-545.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31-95.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance*. Oxford: Basil Blackwell.
- Sperber, D., & Wilson, D. (1995). "Postface" to the second edition of *Relevance*, *Relevance*. Oxford: Blackwell.
- Spohn, W. (1993). Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein? In L. Eckensberger & U. Gähde (Eds.), *Ethik und Empirie. Zum Zusammenspiel von begrifflicher Analyse und erfahrungswissenschaftlicher Forschung in der Ethik* (pp. 151-196). Frankfurt a.M.: Suhrkamp.
- Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz & G. Jäger & R. Van Rooij (Eds.), *Game Theory and Pragmatics* (pp. 82-101): Palgrave MacMillan.
- Stampe, D. (1973). *A Dissertation on Natural Phonology*. Bloomington.: Distributed 1979 by Indiana University Linguistics Club.
- van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9, 333-377.
- van Hout, A. (2007). Optimal and non-optimal interpretations in the acquisition of Dutch past tenses. *Proceedings of GALANA*, 2, 159-170.

- Van Rooy, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, 27, 493-527.
- Wilson, D., & Matsui, T. (1998). Recent approaches to bridging: Truth, coherence, relevance. *UCL Working Papers in Linguistics 10*
- Zeevat, H. (2000). The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17, 243-262.
- Zeevat, H. (2002). Explaining presupposition triggers. In K. van Deemter & R. Kibble (Eds.), *Information Sharing* (pp. 61-87). Stanford: CSLI Publications.
- Zeevat, H. (2004). Presupposition Triggers, Context Markers, or Speech Act Markers. In R. Blutner & H. Zeevat (Eds.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Zeevat, H. (2007a). A Full Solution to the Projection Problem for Presuppositions: University of Amsterdam.
- Zeevat, H. (2007b). *Optimal Interpretation as an Alternative to Gricean Pragmatics* Unpublished manuscript, Universiteit van Amsterdam.
- Zwarts, J. (2006). Om en rond: Een semantische vergelijking. *Nederlandse Taalkunde*, 11, 101-123.

Optimality-Theoretic Pragmatics Meets Experimental Pragmatics

Reinhard Blutner

ILLC, University of Amsterdam

The main concern of this article is to discuss some recent findings concerning the psychological reality of optimality-theoretic pragmatics and its central part – bidirectional optimization. A present challenge is to close the gap between experimental pragmatics and neo-Gricean theories of pragmatics. I claim that OT pragmatics helps to overcome this gap, in particular in connection with the discussion of asymmetries between natural language comprehension and production. The theoretical debate will be concentrated on two different ways of interpreting bidirection: first, bidirectional optimization as a psychologically realistic online mechanism; second, bidirectional optimization as an offline phenomenon of fossilizing optimal form-meaning pairs. It will be argued that neither of these extreme views fits completely with the empirical data when taken *per se*.

1 Introduction

Recent approaches to experimental pragmatics (e.g. Noveck & Sperber, 2005) are mainly concentrated on the investigation of scalar implicatures. Characteristically, the interpretive perspective (hearer's view) is taken in this research. A theoretical main issue is to decide between two rivalling theories: Sperber & Wilson's (1986/1995) Relevance Theory (RT) and Levinson's (2000) theory of presumptive meanings or generalized conversational implicatures (GCIs). Levinson claims that GCIs are calculated automatically – i.e. without demanding much processing resources. In contrast, RT argues that the calculation is controlled and is strongly influenced by the available processing resources. Neo-Griceans (Atlas & Levinson, 1981; Horn, 1984; Blutner, 1998; e.g. Atlas, 2005; Horn, 2005) are normally ignored in this research. A defense for this pretermission is that neo-Gricean theories are normative theories that do not directly make predictions about processing. Unfortunately, this argument

exaggerates the philosophical issue of distinguishing between the normative and the naturalistic realm. Surely, these two aspects of understanding human actions can be clearly separated from each other. However, that does not mean they predict different action patterns in most cases. The idea of a rational world isn't so irrational to be excluded in 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 experimental pragmatics as a scientific, explanatory theory of natural language interpretation, there is no deep empirical conflict between interpretation oriented pragmatics and speaker ethics. It seems the speaker better be cooperative (or pretend to be cooperative) if she wants to use language to bring about effects in hearers.

The aim of this article is to close the gap between experimental pragmatics and neo-Gricean theories of pragmatics. The version of neo-Gricean pragmatics I will consider here is called optimality-theoretic (OT) pragmatics. While the automatic/controlled issue of processing has dominated the recent theoretical debate, OT pragmatics will raise several additional issues. One new issue concerns the asymmetries between comprehension and production. How to explain the experimentally observed asymmetries and what is their status in theories of language acquisition? Seeing comprehension and production as different optimization processes, a further research topic concerns the question of how the two optimization processes are integrated with each other (bidirectional optimization). That relates to the psychological reality of bidirectional optimization in the domain of pragmatics. Another new issue concerns the nature of conventionalization (or fossilization) in pragmatics.

The following quotation from Noveck & Sperber (2005) fully applies to the raised new pragmatic issues.

Properly devised experimental evidence can be highly pertinent to the discussion of pragmatic issues, and pragmatics might greatly benefit from becoming familiar with relevant experimental work and from contributing to it.
(Noveck & Sperber 2005, p. 210)

Without careful experimental research linguistic pragmatics cannot really mature and will remain in a phase of rampant speculation and questionable research habits.

Optimality theory (OT) will be used in this article both in the *broad sense* of a general methodology dealing with resolving conflicting constraints by using universal optimization procedures and in the *narrower sense* of developing an explicit model that concern the essentials of neo-Gricean pragmatics.

In the following sections we assume some familiarity with the basic conceptions of OT pragmatics as provided in the first paper of the present collection. Further, I assume some knowledge about the three main views conforming to a naturalistic pragmatics: RT, Levinson's (2000) theory of presumptive meanings, and the neo-Gricean approach. In the first contribution to this volume, I have demonstrated how the idea of optimal interpretation can be used to restructure the core ideas of these three different approaches. Section 2 explains the idea of fossilization. It is pointed out how the general setting of cultural evolution can help to make this idea precise. Further, a series of important theoretical problems is raised - mainly concerning the distribution of labor between online processes (optimization procedures) and offline processing (fossilization processes). In section 3, I discuss several experimental findings and come to a preliminary conclusion about the relationship between online processes and fossilization phenomena. Section 4 draws some general conclusions relating to a deeper understanding of the idea of naturalization and (cultural) embodiment in the context of natural language interpretation.

2 Fossilization: a bidirectional OT account

In the first contribution to this collection, I have introduced weak bidirectionality and it was illustrated how this solution concept explains Horn's division of pragmatic labour. If we assume that the optimization procedure is supplemented by a system of ranked (heuristic) constraints – in order to provide the content of the optimization – then Horn's R-principle/Q-principle is in exact correspondence to interpretive/expressive optimization. Further, the modulo-clause in the formulation of the Q-/R-principle is explicitly expressed by the recursive term in formalism defining weak bidirectionality.

An important question concerns the *status* of the theory with regard to synchrony versus diachrony. Obviously, both RT and Levinson's theory of presumptive meanings take the synchronic view and both suggest a model of online language interpretation. Within the neo-Gricean camp, the situation is not so clearly decided. Whereas researchers like Atlas (2005) take a synchronous view, researchers like Horn (1984) clearly emphasize the diachronic perspective.

In the framework of OT pragmatics it is very natural to take weak bidirection 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 time scale. Hendriks et al. (to appear) put this point as follows:

On Blutner and Zeevat's evolutionary view of bidirectionality, form-meaning pairs that have been determined by bidirectional optimization constitute fixed relations to a learner who sets out acquiring the language. No learner, indeed no user of the language, needs to perform a bidirectional computation for any form-meaning pair she encounters.

In contrast to this view there are representatives of OT pragmatics who suggest a procedural formulation of weak bidirectionality and propose it as a realistic model of natural language interpretation and/or natural language production (e.g. Zeevat, 2000; Jäger, 2002; Beaver & Lee, 2004; Hendriks & Spence, 2005/2006). This position is also taken in Hendriks et al. (to appear):

However, we take the position that bidirectionality is not in the first place an evolutionary mechanism. Some form-meaning pairings have not been fossilized or automatized, but must be computed anew in a given situation. This view of bidirectionality raises the question of whether bidirectionality is a property of an individual's linguistic performance from the onset of language acquisition, or whether it is acquired or instantiated at some later time. We believe that the latter is the case. Whenever a bidirectional pair has to be computed online in a given situation, it is necessary for the hearer to realize which options were available to the speaker, and also to realize that the speaker's eventual choice is codetermined by the speaker's assumption that the hearer is able to share his perspective. It is to be expected that such online computation requires considerable cognitive resources.

In section 3, I will discuss recent empirical studies that relate to the two different positions.

In natural language pragmatics, the idea of fossilization was introduced first in Geis & Zwicky's (1971) paper about 'invited inferences' as a mechanism of conventionalization for implicatures. A closely related approach is Morgan's (1978) theory of *short-circuited implicatures* where some fundamentally pragmatic mechanism has become partially grammaticalized. Leaning on this idea, Horn & Bayer (1984) propose an elegant account of so-called neg-raising, "the availability (with certain predicates) of lower clause understandings for higher clause negations" (p. 397). There is a principal difficulty for nonsyntactic treatments of these neg-raising interpretations. The difficulty has to do with the existence of lexical exceptions to neg-raising, i.e. we find pairs of virtual synonyms of which one member allows the lower clause understanding and the

other blocks it.¹ Horn & Bayer (1984) argue that conversational implicatures may become conventionalized (“pragmatic conventions”) and this conventionalization sanctions neg-raising. The short-circuiting of implicatures as a matter of convention has important empirical consequences, some of them we will discuss in the following section.²

In an early paper, Cole (1975) investigates similar phenomena in the lexical realm. Calling the conventionalization phenomenon “lexicalization of contextual meaning” he makes quite clear that the relevant conventions are built on the basis of particularized conversational implicatures (i.e. what Levinson (2000) calls *utterance token meanings*). Further, he proposes a diagnostics for discriminating between implicatures proper and their lexicalized counterpart. This may help to clarify the synchronic/diachronic status of conversational implicatures.

Traugott and her colleagues (e.g. Traugott, 1989; Traugott & Dasher, 2005) applied the idea of fossilization to explain language change. According to this model innovation may arise in the individual and spread or propagate through the community. In their *invited inferencing theory of semantic change*, Traugott and co. postulate a cycle starting with coded meaning, exploiting particularized conversational implicatures, transforming these implicatures into generalized conversational implicatures (= conventionalization), and finally resulting in new coded meanings (cf. Traugott & Dasher, 2005). Figure 1 shows a simplified picture of this model.

¹ One of Horn & Bayer's (1984) examples concerns opinion verbs. For instance, Hebrew *xogev* 'think' permits NR readings while *maamin* 'believe' does not. Interestingly, the opposite pattern obtains in Malagasy. In French *souhaiter* 'hope, wish' exhibits neg-raising, but its near-synonym *espérer* does not – although it's Latin etymon *sperare* did. (cited after Horn & Bayer, 1984, p. 400).

² For example, we expect to find differences between speakers and between languages as to just which conventions of usage are operative. And exactly this happens as it is pointed out in Horn & Bayer (1984).

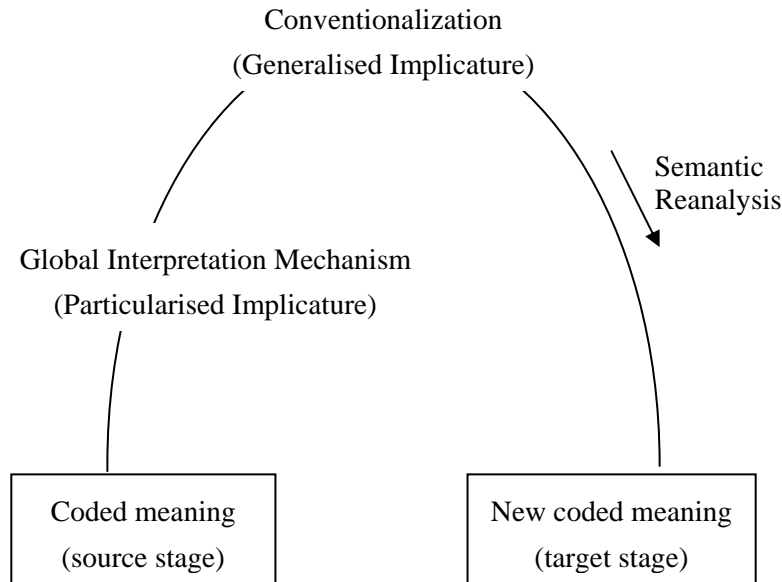


Figure 1: Simplified representation of the *invited inferencing theory of semantic change* (see Traugott & Dasher 2005: 38)

In the domain of syntax, Levinson (2000) und Mattausch (2004) used very the same idea for explaining the development of binding principles.

I will use the term *fossilization* here in a very broad sense that covers the whole spectrum of the mentioned phenomena. It stands for processes of *individual fossilization* or *routinization* that take place in individual language acquisition, i.e. on a time scale of seconds, hours and months. What's more it stands for social processes of *cultural fossilization* that take place in language change on a historical time scale of years up to centuries; the relevant mechanism is iterated learning/cultural evolution.

In OT pragmatics, fossilization relates to a transformation of knowledge systems. As we have seen in the first contribution to this volume, it is possible to describe the same solution space in two different ways. In the first case (Figure 2a, p. 11) unidirectional optimization (either hearer or speaker perspective) is sufficient to calculate the solution pairs. It is plausible to assume that this kind of OT systems can be used to construct cognitively realistic models of online, incremental interpretation (see Blutner, 2006, 2007). The second case (Figure 2b, p. 11) is using the recursion of weak bidirection (super-optimality) and has a completely different status. Because of its strictly non-local nature the proposed

algorithms that calculate the super-optimal solutions do not even fit the simplest requirements of psychologically realistic models of online, incremental interpretation (Zeevat, 2000; Beaver & Lee, 2004).³

The proper understanding of weak bidirection and super-optimality relates best to an off-line mechanism that is based on bidirectional learning (Blutner, Borra, Lentz, Uijlings, & Zevenhuijzen, 2002; Benz, 2003; Van Rooy, 2004; Benz, 2006). In these approaches the solution concept of weak bidirection is considered as a principle describing the results of language change: super-optimal pairs emerge over time in language change. This relates to the view of Horn (1984) who considers the Q and the I principle as diametrically opposed forces in language change, and it conforms to the good old idea that synchronic structure is significantly informed by diachronic forces.

For the sake of illustration let's go back to the example illustrated in Figure 2 of the first contribution to this collection (p. 12). Let's assume a population of agents who realize speaker- and hearer strategies based exclusively on the markedness constraints F and M. In this population each content is expressed in the simplest way (f_I) and each expression is understood in the simplest way (m_I). Let's assume further that these agents communicate with each other. When agent x is in the speaker role and intends to express m_I , then expressive optimization yields f_I . Agent y is a hearer who receives f_I and, according to interpretive optimization, he gets the interpretation m_I – hence the hearer understands what the speaker intends: successful communication. Now assume the speaker wants to express m_2 . With the same logic of optimization he will produce f_I and the agent y interprets it as m_I . In this case, obviously, the communication is not successful. Now assume some kind of *adaptation* either by iterated learning or by some mutations of the ranked constraint system (including the linking constraints). According to this adaptation mechanism the expected 'utility' (how well they understand each other in the statistical mean) can improve in time. In that way a system that is evolving in time can be described including its special attractor dynamics. In each case there is a stabilizing final state that corresponds to the system of Figure 2a (p. 12) where the two Levinsonian (2000) constraints **I** (= [F→M]) and **M** (= [F→M]) outrank the rest of the constraints. It is precisely this system that reflects Horn's division

³ There are several arguments why bidirectional OT cannot yield an online mechanism of linguistic competence. Beaver & Lee (2004) argue that if more rounds of optimization are allowed, the bidirectional OT-model severely overgenerates in the sense that in later rounds peculiar new form-meaning pairs will emerge as winners. Before the Beaver & Lee paper, Zeevat (2000) argued against the symmetric view of OT pragmatics starting from the famous rat/rad problem and its pragmatic counterparts.

of pragmatic labour. The only condition we have to assume is that the marked contents are less frequently expressed than the unmarked contents.⁴

Hence, the important insight is that a system that is exclusively based on markedness constraints such as in Figure 2b (p. 12) is evolutionary related to a system based on highly ranked linking constraints such as in Figure 2a. It is opportune to present some more details at this point. Our own simulation studies (Blutner et al., 2002) have provided the following results assuming the three different strategies illustrated in Figure 2. Here the *Horn-strategy* describes the famous pattern of iconicity (Horn's division of pragmatic labour). The *anti-Horn-strategy* describes a kind of anti-iconicity, and the Smolensky-strategy describes the presumed initial state of a learner where unmarked forms and unmarked meanings are preferred simultaneously.

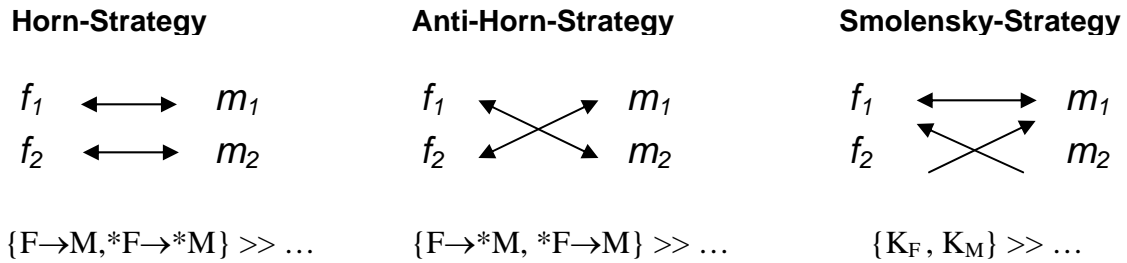


Figure 2: Three different strategies – based on the indicated three different rankings of the constraints

- Horn- und anti-Horn-strategies are the only evolutionary stable strategies.
- If the initial state represents a uniform Smolensky-population, then the systems develops into
 - a pure Horn-population, assumed the frequency of the realization of m_1 is higher than that of m_2 : $P(m_1) > P(m_2)$
 - a pure anti-Horn-population, assumed the frequency of the realization of m_2 is higher than that of m_1 : $P(m_2) > P(m_1)$
- The corresponding proposition is true if the initial state represents a mixed population

⁴ For more discussion of the role of frequencies in an evolutionary setting cf. Stalnaker (2006). The general conclusion is that the solution concept of weak bidirection can be seen as a rough first approximation to the more adequate solution concepts of evolutionary game theory that describe the results of language change.

Hence, the probabilities for the situations that are described, i.e. $P(m_1)$, $P(m_2)$, are decisive for the result. The classical pattern of iconicity is predicted only in cases where the unmarked situation has the highest probability. McCawley (1978) has listed numerous cases of constructional iconicity in the lexicon. Krifka (2007) has observed that the phenomenon is the decisive factor in determining the precise/vague interpretation of measure expressions.

Interestingly, there are also examples of anti-iconicity. They are found in connection with semantic broadening where the initial meaning is described as that of an ideal shape, figure or state. A good example can be found in Dutch, where besides the preposition *om* (= Engl. *round*; German *um*) the expressions *rond* and *rondom* are in use. The expression *rond* is a word borrowed from French. It refers to the ideal shape of a circle. Starting with its appearance it comes in competition with the original (and *unmarked*) expression *om*. The results is a division of labour as demonstrated in the following examples (cf. Zwarts, 2003, 2006):

- (1) a. Ze zaten rond (?om) de televisie
‘They sat round the television’
- b. Een man stak zijn hoofd om (?rond, ?rondom) de deur
‘A man put his head round the door’
- c. De auto reed om (?rond, ?rondom) het obstakel heen
‘The drove round the obstacle’
- d. het gebied rondom (?om) het stadje
‘the area round the little town’

According to the principle of iconicity we would expect that the unmarked form (*om*) is paired with the ideal of the circle shape and the marked form (*rond*) with the detour interpretation.⁵ However, the opposite is true. I think there is a simple explanation for this fact: ideal shapes/situations are much less frequent than non-ideal situations; hence, since $P(m_1) < P(m_2)$, the present evolutionary approach predicts anti-iconicity.

I think these examples and many other examples in the area of lexical pragmatics (e.g. Blutner, 1998; Wilson, 2003) strongly suggest the reality of fossilization. Accepting that both possibilities are real to some extent – the online calculation of implicatures and the access of their fossilized counterparts, the question concerns the distribution of labor between online processes

⁵ The assumption that the ideal path description (circle) is realizing the unmarked interpretation and the detour interpretation is realizing the marked interpretation is justified by independent thoughts about the preference of the logically strongest interpretation (e.g. Dalrymple, Kanazawa, Kim, Mchombo, & Peters, 1998).

(calculating optimal outcomes) and offline processing (fossilization processes). We can ask this question for standard scalar implicatures, as well as for other types of pragmatic inferences. In the next section I will review some experiments that are claimed to decide the issue. These experiments are closely related to the issue of asymmetries between comprehension and production processes.

3 Asymmetries between natural language comprehension and production

It's a common observation that we often are not able to produce what we can understand. The opposite situation, where we are able to produce a certain expression but unable to understand, it is observed much less often. The phenomenon of aphasia gives a feasible illustration of the existence of both kinds of asymmetries (e.g. Jakobson, 1941/1968). Likewise, in the domain of language acquisition both sides of the phenomenon can be detected. It is well known that children's ability in production lags dramatically behind their ability in comprehension (e.g. Benedict, 1979; Clark, 1993). It was only recently that attention was devoted to the opposite case where children's comprehension performance lags years behind their ability of production (cf. Hendriks & Spenader, 2005/2006).

There are three different ways to deal with these observations. The first approach is to assume dissociation between a comprehension grammar and a production grammar. Unfortunately, this account requires some ad hoc stipulations which conflict with general assumptions of parsimony.

The second approach is to assume different processing restrictions for production and comprehension. Joshi (1987) was possibly the first who discussed the asymmetry issue from the viewpoint of artificial intelligence:

Comprehension and generation, when viewed as functions mapping from utterances to meanings and intentions and vice versa, can certainly be regarded as inverses of each other. However, these functions are enormously complex and therefore, although at the global level they are inverses of each other, the inverse transformation (i.e., computation of one function from the other) is not likely to be so direct. So, in this sense, there may be an asymmetry between comprehension and generation even at the theoretical level. (Joshi 1987, p. 183)

Joshi further suggests (p. 184) that the human generation mechanism involves some monitoring of the output, presumably by the comprehension mechanism.

A corresponding monitoring (by generation) is not assumed for the human comprehension mechanism.

The third way of dealing with the asymmetry follows from an optimization approach. This was first demonstrated by Smolensky (1996). As we have seen in the previous sections, natural language production in OT goes from a given interpretation to an optimal expression and natural language comprehension goes from a given expression to an optimal interpretation. It is these different directions of optimization which impose different boundary conditions on the process of optimization. As a result, the same system of constraints and the same constraint hierarchy can account for the observed asymmetry, without taking recourse to multiple grammars or different processing restrictions for production and comprehension.

In this section I will discuss asymmetries between comprehension and production in the context of recent experimentation. The natural language expressions investigated are pronouns, reflexives, referential and quantifying expressions – the latter in connection with scalar implicatures. The fundamental questions asked are twofold:

- (i) How to explain the observed differences between comprehension and production in a certain stage of development?
- (ii) What is the mechanism that handles how to overcome the gap between comprehension and production during natural language acquisition?

OT has a very simple answer to the question (i). In order to account for the usual observation that comprehension can be perfect while production is not, Smolensky (1996) assumes two kinds of constraints: (a) markedness constraints for forms and (b) linking (faithfulness) constraints – linking forms and meanings in an adequate way. Further, he assumes that the markedness constraints initially dominate the linking constraints. It is exactly under these conditions that we get the expected pattern.

For sake of illustration, let us go back to the example with two forms and two meanings (first article of this volume). We introduced the markedness constraint for forms F and the two linking constraints $F \rightarrow M$ and $*F \rightarrow *M$ (see table 1, p. 10). If $\{F\} \gg \{F \rightarrow M, *F \rightarrow *M\}$ then comprehension is always correct (interpreting f_1 as m_1 and f_2 as m_2). However, the production perspective sometimes gives the wrong result. This is because of the dominance of the markedness constraint F , and it gives the result that all meanings m_i ($i = 1, 2$) are expressed by the simpler form f_1 .

Interestingly, the opposite pattern of delayed comprehension is also possible. In this case we have to assume an incomplete system of linking constraints that outranks the system of markedness constraints. A very simple

example is $\{F \rightarrow M\} \gg \{F\}$. In this case m_1 produces f_1 and m_2 produces f_2 . However, while f_1 is always interpreted correctly as m_1 the form f_2 comes out as ambiguous. It can be interpreted both as m_1 and m_2 , and this constitutes a case of delayed comprehension.

The research question (ii) is much more difficult to answer. The difficulty arises from the fact that there is not only one potential mechanism to overcome the gap between comprehension and production. There are at least two such mechanisms, and I will consider them in correspondence with the two ways of viewing bidirection discussed earlier. The first mechanism is based on an OT learning mechanism that re-ranks the involved constraints. That's exactly Smolensky's view as taken in Smolensky (1996). The second mechanism is a mechanism of maturation resulting in a processing system that integrates the comprehension and the production perspective. The resulting integrated system can be either the symmetric system of bidirectional OT or an asymmetric version such as proposed by Joshi and worked out by Zeevat (2000).

In a slightly different formulation, the first mechanism is realizing the diachronic view of bidirection where bidirectional optimization takes place offline (during language acquisition) and leads to some kind of fossilizing optimal form-meaning pairs. In contrast, the second mechanism presumes bidirectional optimization as a psychologically realistic online mechanism. According to this online/synchronic view, speakers (hearers) optimize bidirectionally and take into account hearers (speakers) when selecting (interpreting) a referring expression. In the following I will consider some experimental investigations that shed a light on the empirical adequacy of these two positions.

3.1 The Pronoun Interpretation Problem

In a recent research article Hendriks & Spenader (2005/2006) give a new interpretation of children's delay of the comprehension of pronouns (see also Hendriks, Rijn, & Valkenier, 2007). I discuss the validity of their interpretation and present an alternative account in terms of iterated learning.

A series of experiments has shown that children make errors in interpreting pronouns as late as age 6;6, yet correctly comprehend reflexives from the age of 3;0 (e.g. Chien & Wexler, 1990; McKee, 1992; Koster, 1993; Spenader, Smits, & Hendriks, 2007). For example, children were confronted with a context where two boys, Bert and Paul, are introduced, and the following sentences were given:

- (2) a. Bert is washing himself
- b. Bert is washing him

Sentences like (2a) are correctly understood from a young age (95% of the time according to some studies). However, children misinterpret the pronoun in (2b) as coreferring with the subject about half the time. Hence, it seems that children did not yet realize that the coreferring reading of (5b) must be blocked given the existence of the sentence (2a) which clearly has the coreferring reading.

Contrasting with the comprehension data, language production experiments consistently have shown that children do not have problems in producing reflexives or pronouns correctly. For example, Bloom et al. (1994) demonstrated that even in the youngest age groups investigated (ranging from 2;3 to 3;10) the children consistently used the pronoun to express a disjoint meaning, while they used the reflexive to express a coreferential interpretation. It can be concluded from the production data that children have competence of binding principles. Why they don't use this knowledge in comprehension?

I cannot go into all the different theoretical proposals concerning the pronoun interpretation problem. Instead, I will be mainly concentrated on the possibilities opened by OT pragmatics. Recently, several authors have argued that the observed delay in comprehension can be explained by assuming that children are only able to consider their own perspective, whereas adult hearers are able to simultaneously take into account the perspective of the speaker (deHoop & Kramer, 2005/2006; Hendriks & Spennader, 2005/2006; Hendriks, Rijn et al., 2007).

As explained at the beginning of this section it is possible to account for the delay of comprehension by assuming an incomplete system of linking constraints that outranks the system of markedness constraints for forms, for instance the system $\{F \rightarrow M\} \gg \{F\}$. In the concrete case of pronoun/reflexive interpretation f_1 stands for the reflexive, f_2 for the pronoun, m_1 for the coreferential interpretation and m_2 for the disjoint interpretation. The markedness constraint F prefers the reflexive over the pronoun and can be read as "referential economy" (see Burzio, 1998). The linking constraint $F \rightarrow M$ excludes the reflexive from the disjoint interpretation – that's just the binding principle A (a reflexive must be bound locally) expressed as a violable constraint. Figure 3 shows the preferences between the four possible form meaning pairs arising from the system.

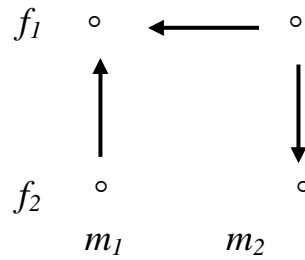


Figure 3: Diagram illustrating the pronoun interpretation problem. It shows the preferences between the four form-interpretation pairs based on the system $\{F \rightarrow M\} \gg \{F\}$ of ranked constraints

Assume now that children begin with unidirectional optimization. In the case of production everything goes right: the meanings m_1 and m_2 are expressed by f_1 and f_2 , respectively. However, in case of comprehension the form f_2 (the pronoun) exhibits an ambiguity: both the interpretation m_1 and m_2 are optimal taken the interpretive perspective for optimization. And that's exactly the expression of the pronoun interpretation problem.

Optimizing bidirectionally inherently involves reasoning about alternatives not present in the current situation. In the present case a child who is hearing f_2 (a pronoun) must reason what other non-expressed forms the speaker could have used. It can realize then that a coreferential meaning m_1 is better expressed with f_1 (a reflexive). Then, by a process of elimination, the child must realize the pronoun should be interpreted as disjoint meaning m_2 and this resolves the ambiguity. Since the ability to optimize bidirectionally may be a skill acquired relatively late, this idea gives a plausible explanation of the lag in acquisition.

Summarizing, the online processing account of Hendriks & Spenader (2005/2006) provides a new way to explain children's delay of the comprehension of pronouns. What's essential for this solution is the hypothesis that the hearer has to take a potential speaker into account. Thus, the authors are able to derive principle B effects (pronouns are free) from principle A alone, through bidirectional optimization. The approach nicely combines a pragmatic explanation with a processing account (lack of processing resources). Besides the stipulation of the constraints and their ranking no other stipulations are required.

However, there are also some arguments that challenge the discussed view. First at all there is the question of constraint grounding. Other systems of

constraints are conceivable and successfully used in the literature (see, e.g., Levinson, 2000; Mattausch, 2004). Further, there is no answer on the question why the particular ranking $\{F \rightarrow M\} \gg \{F\}$ is assumed. Another problem has to do with children's abilities for mind reading (*theory of mind*) that is explicitly assumed in Hendriks' and Spenader's approach. The assumption of mind reading as a prerequisite for making the transition to bidirectional reasoning has the consequence that there should be strict correlations between the behaviour in standard tests of theories of mind (see Perner, Leekam, & Wimmer, 1987) and the behaviour in tasks involving bidirectionality (such as pronoun interpretation). Unfortunately, such strict correlations never were found (Flobbe, Verbrugge, Hendriks, & Krämer, 2007). Further, mind reading requires awareness of other conversation participant's choices. Hence, it is based on controlled rather than automatic processing. However, pronoun processing appears to be automatic rather than controlled. There is no explicit hint for mind reading capacities in such tasks.

In the following subsection I will propose an alternative account that can describe the same kind of data and in addition has some conceptual advantages.

3.2 Pronoun interpretation and related task: individual fossilization

In section 2, I described an approach to fossilization and I made a distinction between individual fossilization (or routinization) and cultural fossilization. Cultural fossilization was successfully used by Mattausch and Jäger (Jäger, 2004; Mattausch, 2004). I will consider now individual fossilization in connection with the pronoun interpretation problem.

In the informal description given here the focus is on pointing out the differences to the processing account provided by Hendriks & co. Let's start with Hendrik's initial system $\{F \rightarrow M\} \gg \{F\}$. In order to apply OT learning theory we assume that a complete system of constraints is present in a background of equally ranked constraints. The following system which is functionally equivalent with the system described before is used: $\{F \rightarrow M\} \gg \{F\} \gg \{F \rightarrow *M, *F \rightarrow M, *F \rightarrow *M\}$. The learning rule then says: promote constraints that favour wanted behaviour over unwanted, demote constraints that favour unwanted behaviour over wanted. If a competent adult acts as speaker and the child as hearer, then this learning rules lead to the promotion of $*F \rightarrow *M$ (principle B). Figure 4 illustrates the transfer between the two systems.⁶

⁶ Alternatively, we could start with the system $\{*F, *M\} \gg \{F \rightarrow M, F \rightarrow *M, *F \rightarrow M, *F \rightarrow *M\}$. The two dominating constraint $*F$ and $*M$ express that f_2 (pronoun) is the preferred form and that m_2 (disjoint interpretation) is the preferred interpretation. The linking constraints cancel each other. Then it can be shown that iterated learning leads to different stages of development. First the principle A is evolving if the plausible

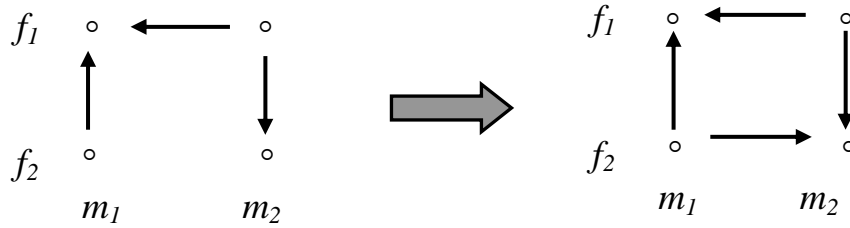


Figure 4: Transformation between two systems of ranked constraints provided by individual fossilization

It is obvious that this transfer is dependent of parameters of use. Hence, we expect frequency effects when the fossilization mechanism is at work. Further, we would expect no significant differences between the comprehension of pronouns and the comprehension of reflexives. The reason is that their processing loads are not significantly different. Hendriks' online view of processing (involving bidirection) conflicts with both hypotheses. It suggests a domain-independent transition from the unidirectional to the bidirectional case. Consequently, we shouldn't expect significant effects of use (frequency effects). Further, for adult subjects we should expect significant differences in processing between pronouns and reflexives, since the pronoun requires bidirectional processing but the reflexive does not.

I think both hypotheses supporting the fossilization view can be confirmed. Though there is no direct verification of the second hypothesis at the moment, I think in the light of the eye tracking investigations of Karabanov, Bosch, & König (to appear) it is not probable that the comprehension of pronouns takes significantly more time than the comprehension of reflexives assuming comparable conditions. For the first hypothesis, it's important to see that there are some other domains which realize the same structural relations as exhibited in the case of pronoun interpretation shown in Figure 3. Consider first the domain of natural language quantifiers and consider dual quantifiers such as *some*(A) and *all*(A), where A stand for a certain restrictive term. Logically, *all*(A)(B) has the set inclusion interpretation stating $A \subseteq B$, and *some*(A)(B) has an interpretation expressing nonempty intersection $A \cap B \neq \emptyset$. Of course, this

stipulation is made that $P(m_2) > P(m_1)$. Hence we have a motivation why the system of preferences as given on the left hand site of Figure 4 appears – it reflects delayed comprehension – instead of a system exhibiting delayed production. Only later the principle B becomes dominant, giving the preferences shown on the right hand site of Figure 4.

interpretation does not exclude the set inclusion reading. It's the scalar implicature that excludes this interpretation – leading to a *some_but_not_all* interpretation. The ordering of all form interpretation pairs given in Figure 3 can be applied to the quantifier case when we assume that f_1 stands for *all*, f_2 for *some*, m_1 for the set inclusion interpretation, and m_2 for the *some_but_not_all* reading. The markedness constraint F now prefers *all* over *some*. We can see that as a realization of the strongest meaning hypothesis (Dalrymple et al., 1998). Further, the dominating constraint $F \rightarrow M$ expresses the meaning postulate for *all*, and the potential constraint $*F \rightarrow *M$ expresses the scalar implicature for *some*.

The first systematic investigation of the acquisition of scalar implicature can be attributed to Noveck (2001). From his experiments it can be concluded that young children initially treat a relatively weak term logically before becoming aware of its pragmatic potential, and that, in this respect, “children are more logical than adults” (Noveck, 2001: 165). Concluding, we can speak of delayed comprehension of the pragmatic potential of the weak quantifier.

Another domain where we find similar effects is the interpretation indefinite expressions. In several languages it has been observed that indefinite noun phrases such as *a boy* take on different interpretations depending on whether they appear in a scrambled or unscrambled word ordering (e.g. de Hoop & Krämer, 2005; Unsworth, 2005). Adults interpret unscrambled indefinites (f_1) as ‘non-specific’ (m_1) whereas they interpret scrambled indefinites (f_2) as ‘specific’ (m_2). Again we find a delayed comprehension effect: children interpret scrambled indefinites in both ways. Only later they realize that the ‘specific’ interpretation is the proper one.

In a recent article Hendriks et al. (2007) discuss the results of diverse experiments in different domains and conclude that children seem to differ in the ages at which they provide adult-like responses for particular linguistic forms.

Whereas from the age of 6 or 7 on children start to interpret pronouns correctly, children until roughly 11 years old select a non-adult meaning for indefinite objects (Unsworth, 2005), and many 10- and 11-year-olds do not draw a scalar implicature where most adults would (Noveck, 2001). This suggests that bi-directional optimization is not a general strategy that has to be learned by children in one step, but rather that the possibility of bi-directional optimization is dependent on the frequency of use of the relevant production rules. (p. 1893)

Hence, the first hypothesis suggested above – predicting a domain-independent transition from the unidirectional to the bidirectional view – seems to be

falsified. And this might be a powerful argument supporting the fossilization view.

Though the domain independence of the transfer from unidirectional to bidirectional processing is a natural consequence of the online processing view, it is not a necessary consequence. Hendriks et al. (2007) provide an improvement of their online processing view in order to describe the empirically found domain dependency. This improvement is formulated in terms of the ACT-R model (Anderson & Lebiere, 1998; Anderson et al., 2004).

ACT-R understands itself as an integrated theory of the mind. Different from Smolensky's (Smolensky & Legendre, 2006) theory of *harmonic mind* which sees the symbolic part (i.e., OT) as a high-level description of the neural realm, ACT-R is a *hybrid* theory that relates different symbolic modules with certain subsymbolic processes. These subsymbolic processes serve to guide the selection of rules to fire as well as the internal operations of modules and much of learning.

Hendriks et al. (2007) model unidirectional and bidirectional OT in terms of the ACT-R model. In this model bidirectional optimization is described as the serial application of two unidirectional processes of optimization. A crucial property of ACT-R is the assumption that actions take time to perform and that performance is limited by the serial processing bottleneck. Since bidirectional optimization needs much more processing resources than unidirectional optimization does, a process of production compilation⁷ comes in increasing the processing efficiency. The result of product compilation conforms to an instance based kind of automatization (Logan, 1988). I think what is described here comes very close to the idea of fossilization. Whereas fossilization leads to the introduction of new linking constraints product compilation leads to the generation of new productions who describe the results of certain bidirectional actions.

3.3 Choosing the right referring expression

The standard case of production/comprehension asymmetries is delayed production. Comprehension can be perfect while production is not. A good example is given by production and understanding of R-expressions and pronouns as illustrated in (3).

⁷ In production compilation, two existing production rules are integrated into one new production rule. Production compilation occurs when two existing production rules are repetitively executed in sequence.

- (3) Discourse context: A woman is waiting at the corner. Her girl is eating an ice cream cone.
- a. She wears a red shirt.
 - b. The woman wears a red shirt.

The interpretation of the pronoun in (3a) clearly refers to the discourse topic (*the girl*). If we want to express the alternative meaning as in (3b) we cannot use the pronoun. Interestingly, young children very often produce such subject pronouns when intending to refer to non-topics. Karmiloff-Smith (1985) found this pattern of production in children until the age of 6.

I have already mentioned that the phenomenon can be modeled by assuming markedness conventions that initially dominate linking constraints. Figure 5 shows the corresponding diagram. Hereby, f_1 stands for the pronoun and f_2 for an R-expression. Further, m_1 is the interpretation referring to the topicalized discourse referent while m_2 refers to the non-topicalized one. F can be seen as referential economy (preferring pronouns to R-expressions) and $F \rightarrow M$ expresses the preference for pronouns to be interpreted as the topic of the discourse.

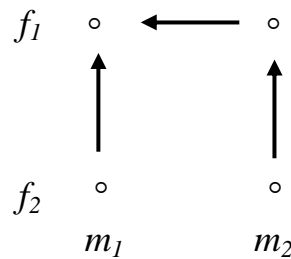


Figure 5: Diagram illustrating the pronoun production problem. It shows the preferences between the four form-interpretation pairs based on the system $\{F\} \gg \{F \rightarrow M\}$ of ranked constraints

Using unidirectional optimization, the diagram describes the OT system of an agent who can properly understand pronouns and R-expressions but who overuse pronouns when intending to refer to non-topics.

The two considered models now make different assumptions for describing the transfer from the child system to the adult system. The online processing model handles the production problem by assuming that the producer takes the hearer into account and begins to reason bidirectionally at some point of development. In contrast, the fossilization view says that unidirectional

optimization is sufficient if it is assumed that the relevant information has been fossilized at some part of the human development.

In a recent research article, Hendriks, Englert, & Wubs (2007) argue that the investigation of elderly adults could decide between the two models. Elderly adults possess the required pragmatic and grammatical knowledge to select and interpret referring. However, their linguistic performance can be defective, due to a decreasing working memory capacity. And indeed, the authors found that elderly adults produce non-recoverable pronouns significantly more often than young adults when referring to the old topic in the presence of a new topic. With respect to the comprehension task, no significant differences were found between elderly and young adults.

Obviously, this experimental outcome is a great problem for the fossilization view, since a stipulation of a mechanism of ‘de-fossilization’ does not make any sense in the present context. Hence, the assumption that the speaker takes the hearer into account is well motivated for such examples. Zeevat (2000) has argued for this kind of active, creative processes.

However, there is also a problem for the bidirectional processing view. It says that both the speaker takes the hearer into account and, vice versa, the hearer takes the speaker into account. If that is right, then the same argumentation that is given in the paper by Hendriks, Englert, & Wubs (2007) can be applied for the delayed comprehension experiments discussed in the previous subsections. Thought I don’t know of any experiments with elderly people concerning the delayed comprehension task, I bet more than my finger that the behavior of elderly people does not go down to that of young children in the relevant respects. Hence what we can conclude from these experiments is an asymmetry of processing: the speaker takes the hearer into account but not necessarily vice versa. This is actually Zeevat’s (2000) view of making a distinction between the active and creative process of production and the rather passive process of interpretation.⁸ The idea of fossilization is needed in order to account for the delayed comprehension data.

⁸ “The situation can be fruitfully compared to the habit of hiding easter eggs for one’s children. The parents engaged in hiding the eggs balance the amount of effort with the desired amount of difficulty in finding the egg. (They also picture the child looking for it and try to keep it possible for the child of finding the egg, without spoiling the fun.) For the child it is another matter. They just have to throw in the effort required for finding the eggs. Not more of course, but definitely not less. It is not a complicated balancing act.” (Zeevat 2000: 245)

4 Conclusions

The aim of this article was to close the gap between experimental pragmatics and neo-Gricean theories of pragmatics as formulated in OT pragmatics. I have argued that OT pragmatics has the potential to account both for the synchronic and the diachronic perspective in pragmatics. I further have pointed out that the concept of fossilization can help to understand the idea of naturalization and (cultural) embodiment in the context of natural language interpretation.

Concerning modern pragmatic theories such as RT and Levinson's theory of presumptive meanings, the conflict between effort minimization and effect maximization is resolved in different ways. In a certain sense, the crux of both approaches can be translated in OT pragmatics by making use of particular linking constraints. This translation makes the advantage of both approaches visible: both conform to the incremental, online character of natural language interpretation.⁹

In the last part of the paper I have discussed recent work about the phenomenon of delayed comprehension and delayed production. This is a phenomenon which was not discussed within experimental pragmatics, though the importance of the problem was clearly recognized within OT pragmatics. I have discussed two models which conceptualized bidirection in different ways: the online processing model and the fossilization account. I have argued that neither of these extreme views gives a complete fit to the empirical data when taken per se. While it is obvious that fossilization phenomena are real to some extent it can be argued that a restricted online version of bidirection is correct: speakers optimize bidirectionally and take the hearer into account when calculating the optimal expression; in contrast, hearers normally do not take the speaker into account when calculating the optimal interpretation.

5 References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111, 1036-1060.

⁹ In discussing processing characteristics, incrementality and automaticity of processing have to be discriminated. Whereas automaticity of processing implicates the incremental character of processing the opposite is not true: incrementality does not implicate automatic processing. RT explains the incremental character of processing and has good reasons for assuming controlled processing in order to account for the processing of conversational implicatures. That's different from Levinson's account which assumes automatic processing for generalized conversational implicatures. It seems that RT is better justified on empirical grounds (cf. Noveck & Sperber, 2005).

- Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ.: Erlbaum.
- Atlas, J. D. (2005). *Logic, Meaning, and Conversation: Semantical Underdeterminacy, Implicature, and Their Interface*. Oxford: Oxford University Press.
- Atlas, J. D., & Levinson, S. C. (1981). It-clefts, informativeness and logical form. In P. Cole (Ed.), *Radical Pragmatics* (pp. 1-61). New York: Academic Press.
- Axelrod, R. (1984). *The evolution of co-operation*. London: Penguin.
- Beaver, D., & Lee, H. (2004). Input-output mismatches in OT. In Palgrave/Macmillan (Ed.), *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire.
- Benedict, H. (1979). Early Lexical Development: Comprehension and Production. *Journal of Child Language*, 6, 183-200.
- Benz, A. (2003). Partial Blocking, associative learning, and the principle of weak optimality. In J. Spenader & A. Eriksson & Ö. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory* (pp. 150-159). Stockholm.
- Benz, A. (2006). Partial blocking and associative learning. *Linguistics and Philosophy*, 29, 587-615.
- Biro, T. (2006). *Finding the right words: Implementing Optimality Theory with simulated annealing*. University of Groningen, Groningen.
- Bloom, P., Barss, A., Nicol, J., & Conway, L. (1994). Children's knowledge of binding and coreference: Evidence from spontaneous speech. *Language*, 70, 53-71.
- Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, 15, 115-162.
- Blutner, R. (2004). Nonmonotonic inferences and neural networks. *Synthese (Special issue Knowledge, Rationality and Action)*, 142, 143-174.
- Blutner, R. (2006). Embedded implicatures and optimality theoretic pragmatics. In Torgim Solstad & A. Grønn & D. Haug (Eds.), *A Festschrift for Kjell Johan Sæbø: in partial fulfilment of the requirements for the celebration of his 50th birthday*. Oslo.
- Blutner, R. (2007). Optimality Theoretic Pragmatics and the Explicature/Implicature Distinction. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 67-89). Houndmills, Basingstoke, Hampshire: Palgrave/MacMillan.
- Blutner, R., Borra, E., Lentz, T., Uijlings, A., & Zevenhuijzen, R. (2002). Signalling games: Hoe evolutie optimale strategieën selecteert, *Handelingen van de 24ste Nederlands-Vlaamse Filosofiedag*. Amsterdam: Universiteit van Amsterdam.
- Burzio, L. (1998). Anaphora and soft constraints. In P. Barbosa & D. Fox & P. Hagstrom & M. McGinnis & D. Pesetsky (Eds.), *Is the best good enough?* Cambridge, Mass.: The MIT Press.
- Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

- Chien, Y.-C., & Wexler, K. (1990). Children's knowledge of locality conditions on binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition*, 13, 225-295.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond* (pp. 39-103). Oxford: Oxford University Press.
- Clark, E. V. (1993). *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.
- Cole, P. (1975). The synchronic and diachronic status of conversational implicature. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Volume 3: Speech Acts* (pp. 257-288). San Diego, Cal.: Academic Press, Inc.
- Dalrymple, M., Kanazawa, M., Kim, Y., Mchombo, S., & Peters, S. (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21, 159-210.
- de Hoop, H., & Krämer, I. (2005). Children's Optimal Interpretations of Indefinite Subjects and Objects. *Language Acquisition*, 13, 103-123.
- deHoop, H., & Kramer, I. (2005/2006). Children's Optimal Interpretations of Indefinite Subjects and Objects. *Language Acquisition*, 13, 103-123.
- Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2007). *Children's application of Theory of Mind in reasoning and language*. Unpublished manuscript, University of Groningen & Radboud University Nijmegen.
- Fodor, J. (1975). *The Language of Thought*. New York: Thomas Crowell.
- Gazdar, G. (1979). *Pragmatics*. New York: Academic Press.
- Geis, M., & Zwicky, A. (1971). On invited inference. *Linguistic Inquiry*, 2, 561-579.
- Hendriks, P., & de Hoop, H. (2001). Optimality theoretic semantics. *Linguistics and Philosophy*, 24, 1-32.
- Hendriks, P., Englert, C., Wubs, E., & Hoeks, J. (2007). *Age differences in adults' use of referring expressions*. Unpublished manuscript, University of Groningen.
- Hendriks, P., Hoop, H. d., Krämer, I., Swart, H. d., & Zwarts, J. (to appear). *Conflicts in Interpretation*.
- Hendriks, P., Rijn, H. v., & Valkenier, B. (2007). Learning to reason about speakers' alternatives in sentence comprehension: A computational account. *Lingua*, 117, 1879-1896.
- Hendriks, P., & Spenader, J. (2005/2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, 13, 319-348.
- Hobbs, J., & Martin, P. (1987). Local Pragmatics, *Proceedings, International Joint Conference on Artificial Intelligence* (pp. 520-523). Milan.
- Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.

- Horn, L. (1989). *A natural history of negation*. Chicago: Chicago University Press.
- Horn, L. (2004). Implicature. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics*. Oxford: Blackwell.
- Horn, L. (2005). Current issues in neo-Gricean pragmatics. *Intercultural Pragmatics*, 2, 191-204.
- Horn, L., & Bayer, S. (1984). Short-circuited implicature: A negative contribution. *Linguistics and Philosophy*, 7, 397-414.
- Jäger, G. (2002). Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information*, 11, 427-451.
- Jäger, G. (2004). Learning constraint sub-hierarchies. The bidirectional gradual learning Algorithm. In R. Blutner & H. Zeevat (Eds.), *Pragmatics and Optimality Theory*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
- Jakobson, R. (1941/1968). *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.
- Jaszczolt, K. M. (to appear). Semantics and Pragmatics: The Boundary Issue. In C. Maienborn & K. v. Heusinger & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter.
- Joshi, A. K. (1987). Generation - a new frontier of natural language processing? . *Theoretical Issues in Natural Language Processing*, 3, 181-184.
- Karabanov, A., Bosch, P., & König, P. (to appear). Eye Tracking as a tool for investigating the comprehension of referential expressions. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base*. Berlin: De Gruyter.
- Karmiloff-Smith, A. (1985). Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes*, 1, 61-85.
- Koster, C. (1993). *Errors in Anaphora Acquisition*. Unpublished Ph.D. Dissertation, Utrecht University, Utrecht.
- Krifka, M. (1995). The Semantics and Pragmatics of Polarity Items. *Linguistic Analysis*, 25, 209-257.
- Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma & I. Krämer & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 111-126). Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen.
- Levinson, S. (2000). *Presumptive meaning: The theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492-527.
- Mattausch, J. (2004). *On the Optimization & Grammaticalization of Anaphora*. Unpublished Ph.D. Thesis, Humboldt University, Berlin.

- McCawley, J. D. (1978). Conversational implicature and the lexicon. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 245-259). New York: Academic Press.
- McKee, C. (1992). A Comparison of Pronouns and Anaphors in Italian and English Acquisition. *Language Acquisition*, 2.
- Morgan, J. L. (1978). Two Types of Convention in Indirect Speech Acts. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 261-280). New York: Academic Press.
- Noveck, I. (2001). When children are more logical than adults. *Cognition*, 78, 165-188.
- Noveck, I. A., & Sperber, D. (Eds.). (2005). *Experimental Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave MacMillan.
- Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year Olds' Difficulty with False Belief: The Case for a Conceptual Deficit. *British Journal of Developmental Psychology*, 5, 125-137.
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In J. L. McClelland & D. E. Rumelhart & the-PDP-Research-Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1*. Cambridge, MA: MIT Press.
- Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, 23, 361-382.
- Sæbø, K. J. (2004). Optimal interpretations of permission sentences. In D. de Jongh & P. Dekker (Eds.), *Proceedings of the 5th Tbilisi Symposium on Language, Logic and Computation* (pp. 137-144). Amsterdam and Tbilisi.
- Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27, 367-391.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. *Psychological Review*, 84, 1-66.
- Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, 27, 720-731.
- Smolensky, P., & Legendre, G. (2006). *The Harmonic Mind: From neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.
- Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry*, 13, 483-545.
- Spenader, J., Smits, E.-J., & Hendriks, P. (2007). *Coherent discourse solves the Pronoun Interpretation Problem*. Unpublished manuscript, University of Groningen.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31-95.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance*. Oxford: Basil Blackwell.
- Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz & G. Jäger & R. Van Rooij (Eds.), *Game Theory and Pragmatics* (pp. 82-101): Palgrave MacMillan.

- Traugott, E. (1989). On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change. *Language* 65, 31-55.
- Traugott, E., & Dasher, R. B. (2005). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Unsworth, S. (2005). *Child L2, Adult L2, Child L1: Differences and Similarities*. Unpublished PhD Dissertation, Utrecht University.
- Van Rooy, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, 27, 493-527.
- Wilson, D. (2003). Relevance and Lexical Pragmatics. *Italian Journal of Linguistics/Rivista di Linguistica*, 15, 273-291.
- Wilson, D., & Matsui, T. (1998). Recent approaches to bridging: Truth, coherence, relevance. *UCL Working Papers in Linguistics* 10
- Zeevat, H. (2000). The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17, 243-262.
- Zeevat, H. (2002). Explaining presupposition triggers. In K. van Deemter & R. Kibble (Eds.), *Information Sharing* (pp. 61-87). Stanford: CSLI Publications.
- Zeevat, H. (2007). *Optimal Interpretation as an Alternative to Gricean Pragmatics*. Unpublished manuscript, Universiteit van Amsterdam.
- Zwarts, J. (2003). Lexical Competition: 'Round' in English and Dutch. In P. Dekker & R. van Rooy (Eds.), *Proceedings of the Fourteenth Amsterdam Colloquium* (pp. 229-234). Amsterdam: ILLC.
- Zwarts, J. (2006). Om en rond: Een semantische vergelijking. *Nederlandse Taalkunde*, 11, 101-123.

Neural Networks, Penalty Logic and Optimality Theory

Reinhard Blutner

University of Amsterdam

Ever since the discovery of neural networks, there has been a controversy between two modes of information processing. On the one hand, symbolic systems have proven indispensable for our understanding of higher intelligence, especially when cognitive domains like language and reasoning are examined. On the other hand, it is a matter of fact that intelligence resides in the brain, where computation appears to be organized by numerical and statistical principles and where a parallel distributed architecture is appropriate. The present claim is in line with researchers like Paul Smolensky and Peter Gärdenfors and suggests that this controversy can be resolved by a unified theory of cognition – one that integrates both aspects of cognition and assigns the proper roles to symbolic computation and numerical neural computation.

The overall goal in this contribution is to discuss formal systems that are suitable for grounding the formal basis for such a unified theory. It is suggested that the instruments of modern logic and model theoretic semantics are appropriate for analyzing certain aspects of dynamical systems like inferring and learning in neural networks. Hence, I suggest that an active dialogue between the traditional symbolic approaches to logic, information and language and the connectionist paradigm is possible and fruitful. An essential component of this dialogue refers to Optimality Theory (OT) – taken as a theory that likewise aims to overcome the gap between symbolic and neuronal systems. In the light of the proposed logical analysis notions like recoverability and bidirection are explained, and likewise the problem of founding a strict constraint hierarchy is discussed. Moreover, a claim is made for developing an “embodied” OT closing the gap between symbolic representation and embodied cognition.

1 Introduction

To date, progress in cognitive neuroscience has been hindered by the enormity of the gap between our understanding of some low-level properties of the brain on the one hand, and of some very high-level properties of the mind on the other hand. Research on parallel distributed processing and neural networks (connectionist paradigm) has tried to reduce this gap but was only partially successful. A main characteristic of mainstream connectionism is its *eliminative* character, i.e. the idea that the basic architecture of symbolism (including its

crucial concepts such as representations, rules, compositionality, and modularity) has to be replaced by the concepts of neural networks (cf. Churchland, 1986). In this way, the main advantage of traditional symbolism – the transparency and relative simplicity of descriptions and explanations – are likewise eliminated.

In contrast, there are other researchers who like to play down the neuronal perspective as an issue of implementation. Representatives of this position are, *inter alia*, Fodor and Pylyshyn (1988), who insist that the proper role of connectionism in cognitive science is merely to *implement* existing symbolic theory. According to this view, the *systematicity* of our linguistic competence can be explained only by assuming a classical, symbolist architecture of cognition. If this position reflects an adequate research programme, then the task of overcoming the gap between symbolism and its neural embodiment is not really important for the understanding of our higher-level cognitive abilities.

The methodological position pursued in this article is an *integrative* position. It claims that both modes of computation – symbolic and neural – are theoretically justified and equally important and that there is no need to eliminate one of them. In the case under discussion the point is to assume that symbols and symbol processing are a macro-level description of what is considered as connectionist system at the micro level. This position is not unlike the one taken in theoretical physics, relating, for example, thermodynamics and statistical physics, or, in a slightly different way, Newtonian mechanics and quantum mechanics. Hence, the idea is that the symbolic and the subsymbolic mode of computation can be integrated within a unified theory of cognition. If successful, this theory is able to overcome the gap between the two modes of computation and it assigns the proper roles to symbolic, neural and statistical computation (Balkenius & Gärdenfors, 1991; Smolensky, 1995; Kokinov, 1997; Blutner, 2004; Graben, 2004; Smolensky & Legendre, to appear).

There is a second methodological aspect that relates to the status of theoretical models in integrated research. My primary aim is the demonstration that the tools of logic and algebraic semantics are useful for understanding the emergent properties of neural networks dynamics. However, the dynamics of real neural networks is rather complicated. These systems are perhaps among the most complex known to science. And it is completely unrealistic to understand the emergent properties of such systems by trying to model in detail all what is known about the basic principles of neural operation and causal mechanisms of individual nerve cells. Rather, radical simplification is in order even if these simplifications appear completely unrealistic. These simplifications may lead to different theoretical models which make different views explicit, and this makes it easier to structure the debate for or against a certain position. Theoretical models bring out the hidden assumptions of an approach, particularly with

respect to the elementary neural mechanisms that are required. Moreover, they help to assess the plausibility of certain assumptions, for example with respect to the assumed network architecture. They may invite the construction of new models that make another view and other functional determinants explicit. Even if it is not possible to collect the necessary empirical data to make the model predictions empirically grounded, a lot can still be learned about the causal determinants of certain forms of behavior. Finally, even oversimplified theoretical models may suggest new experiments for empirical data collection.

A third methodological aspect concerns a potential misunderstanding. In the following I will pursue a certain kind of propositional default logic to describe inferences in neural networks. This might suggest that certain logical systems get a deeper justification in terms of neural processing, or it might even suggest that I'm proposing a neural underpinning of certain types of natural reasoning. Hence, it might appear as if we are running in a neuro-cognitive Frege-fallacy by seeing logic as part of cognitive neuroscience. However, such conclusions are unjustified. I only suggest to consider the proposed logical system as a kind of *meta-language* which is useful for modelling certain constraint-based symbolic systems. This is analogous to the use of Prolog as a logical programming language. Without doubt, Prolog can be used for many different applications starting from the modelling of parsing and natural language comprehension and going on to the modelling of planning mechanisms and the abilities of logical inference agents. Nobody would suggest that these applications – if successful – give a deeper justification for Prolog as part of Cognitive Linguistics (at least if we reject the strong view of Artificial Intelligence; see Searle (1980)). In a similar way, the present logical system can be used for many different purposes. This becomes pretty clear when we enlighten the close connection to Optimality Theory (OT) – a general framework which was introduced by Prince & Smolensky (1993/2004) for describing constraint interaction in Generative Grammar.

In the following I will address the issue of formal tools and logical systems which are suitable for grounding the basis for a unified theory of cognition, and I will suggest that an active dialogue between the traditional symbolic approaches to logic, information and language and the connectionist paradigm is possible and fruitful. An essential component of this dialogue refers to OT (Prince & Smolensky, 1993/2004) – taken as a theory that likewise aims to overcome the gap between symbolic and neural systems.

Section 2 introduces symmetric neural networks and explains their basic properties. The idea of inferences in neural networks is explained in Section 3. The developed inferential notion rests on the (non-symbolic) concept of information states and is adequate for describing how neuron activities spread through a symmetric network. Section 4 discusses Penalty Logic – a logic that

was introduced by Pinkas (1995) in order to demonstrate what kind of logical systems symmetric networks can implement. In Section 5 a logic called Penalty/Reward logic is introduced and it is shown that such logic is an adequate tool for dealing with underspecification and conceptual enrichment in symmetric networks. In Section 6 I will discuss the relations to OT, and Section 7 draws some conclusions and shows the connection to recent efforts toward developing an embodied view of cognition.

2 Symmetric networks

Connectionist systems aim at modelling aspects of the nervous system on an abstract computational level. (Good introductions are given in McClelland & Rumelhart, 1986; Rojas, 1996; Bechtel, 2002). The central concept in a connectionist system is the individual unit ('node') which models the functionality of a neuron or a group of neurons. In fact, the units/nodes of most connectionist models are vastly simpler than real neurons. However, such networks can behave with surprising complexity and subtlety. This is because processing is occurring in parallel and usually interactively. In many cases, the way the units are connected is much more important for the behaviour of the complete system than the details of the individual units.

In the following we will assume that the individual units of a connectionist network correspond to larger groups of neurons, sometimes called columns, pools or assemblies (Hebb, 1949; Feldman & Ballard, 1982; Maass, 1999; Wennekers & Palm, 2000). A central idea of the assembly concept is that assemblies can overlap, meaning that one and the same neuron can be part of different assemblies. The organization of assemblies is done according to functional criteria and can be different for different functional contexts. Necessary conditions for constituting an assembly are strong internal couplings within the assembly.

The simplest form of describing the activation dynamics of single units is to assume a nonlinear function that yields the (average) firing rate of the unit given the sum potential of the unit. This sum potential can be calculated by weighted linear combinations of the firing rates of the incoming units. In the present approximation it goes without calculating the full action potentials (spikes). All that is needed are the firing rates of the units, which are directly transferred to the other cells. It has been argued that this method yields a valid approximation of realistic spiking behaviour under certain conditions (for details, see Maass, 1999; Wennekers, 1999; Gerstner & Kistler, 2002). However, it has also been argued that simple rate-based models are not sufficient to model information processing in neuronal systems. There is increasing evidence that the information transferred by a unit consists not only

in the average firing rate but also includes the phase of the spiking functions. This might be relevant for explaining binding by synchronization (e.g. von der Malsburg, 1981; Shastri & Ajjanagadde, 1993; Singer & Gray, 1995). In the following I will simply ignore this complication.¹

There are different kinds of connectionist architectures. In *multilayer perceptrons*, for instance, we have several layers of nodes (typically an input layer, one or more layers of hidden nodes, and an output layer). A fundamental characteristic of these networks is that they are *feedforward* networks, that means that units at level i may not affect the activity of units at levels lower than i . In typical cases there are only connections from level i to level $i+1$. In contrast to feedforward networks, *recurrent networks* allow connections in both directions. A nice property of such networks is that they are able to gather and utilize information about a sequence of activations. Further, some types of recurrent nets can be used for modelling associative memories. If we consider how activation spreads out we find that feedforward networks always stabilize. In contrast, there are some recurrent networks that never stabilize. Rather, they behave as chaotic systems that oscillate between different states of activation.

One particular type of recurrent networks is a *symmetric network*, which is also called a *Hopfield network* (Hopfield 1982). Such networks always stabilize. Hopfield proved that by demonstrating the analogy between this sort of networks and the physical system of spin glasses and by showing that one could calculate a very useful measure of the overall state of the network that was equivalent to the measure of energy in the spin glass system. A Hopfield net tends to move toward a state of equilibrium that is equivalent to a state of lowest energy in a thermodynamic system.

As mentioned already, neural networks can be considered systems of connected units. Each unit has a certain *working range of activity*, which can be represented by an interval $[a, b]$ if an analogous unit is assumed (e.g. Hopfield, 1984; Hopfield & Tank, 1985); a indicates the minimal firing rate of the unit and b indicates the maximal firing rate. Usual choices for the working range of a node are the interval $[0, 1]$ (e.g. Balkenius & Gärdenfors, 1991; Pinkas, 1995) or the interval $[-1, +1]$ (Blutner, 2004). In the latter case the value 0 can be taken as indicating the resting rate. Though neurons with different working ranges can be assumed to be basically equivalent (supposing the thresholds are adapted appropriately), there may be differences (i) due to the interpretation of the activations, (ii) due to the simplicity of the resulting equations, and (iii) due to the stipulation of different discrete subsets when it comes to the introduction of

¹ Some authors doubt that "binding by synchronization" is really such a realistic solution to the binding problem as it often is suggested. For instance, Palm & Wennekers (1997) argue that also other mechanisms are thinkable based on purely rate-based information.

logical values. The discrete values typically taken are $\{0, 1\} \subset [0, 1]$ in the first case (classical binary logic) and $\{-1, 0, +1\} \subset [-1, +1]$ in the second case (tree-valued logic).

A possible state s of the system describes the activities of each node: $s \in [a, b]^n$, with n = the number of units. A possible *configuration* of the network is characterized by a *connection matrix* w . Hopfield networks are defined by symmetric configurations and zero diagonals ($-\infty < w_{ij} < +\infty$, $w_{ij} = w_{ji}$, $w_{ii} = 0$). That means node i has the same effect on node j as node j has on node i , and the nodes don't affect themselves.² The *fast dynamics* describes how node activities spread through that network. In the simplest case this is described by the following *update function*:

$$(1) \quad f(s)_i = \theta \left(\sum_j w_{ij} s_j \right) \quad (\theta \text{ a nonlinear function, typically a step function or a sigmoid function}).$$

Equation (1) describes a nonlinear threshold unit. This activation rule is the same as that of Rosenblatt's perceptron. It is applied many times to each unit. Hopfield (1982) employed an *asynchronous* update procedure in which each unit, at its own randomly determined times, would update its activation (depending on its current net input).³

Using the interval $[0, 1]$ as working range of a unit, Balkenius & Gärdenfors (1991) have argued that the set $S = [0, 1]^n$ of activation states of a network with n units can be partially ordered in accordance with their informational content. Assuming that the vector $\mathbf{0} = \langle 0, 0, \dots, 0 \rangle$ represents a

² It is often mentioned that these assumptions are highly implausible when taking the units of the network as real neurons. It is not clear why real networks should be symmetric and irreflexive. If the assembly idea comes in, we can overcome this problem since it is plausible to assume that the formation of assemblies happens under the pressure of stabilisation, and this might be one of the reasons for symmetry and irreflexivity.

Some people doubt the plausibility of the 'neuron doctrine'. Based on the finding that in the cerebral cortex the majority of neurons have only dendrites and the axons are missing there (this contrasts with the peripheral nervous system where almost every neuron has an axon) (Jibu & Yasue, 1995, p. 100ff). Hence, it has been argued that the working of the cerebral cortex can be better understood by certain microscopic physical processes taking place in the sophisticated network of dendrites of neurons without axons, that is, in the dendritic network (Pribram, 1991; Jibu & Yasue, 1995). The spin-glass model (or, equivalently, the Hopfield network) can be seen as a first approximation to the dendritic network (Jibu & Yasue, 1995). Hence, Hopfield networks can be seen as a good starting point for modelling brain activity independent of whether we accept the neuron doctrine or not.

³ The use of asynchronous updates helps to prevent the network from falling into unstable oscillations.

scheme with minimal informational content and that the vector $\mathbf{1} = \langle 1, 1, \dots, 1 \rangle$ represents maximal informational content, then the following ordering can be seen as reflecting greater *positive* informational content:

$$(2) \quad s \geq t \text{ iff } s_i \geq t_i \geq 0, \text{ for all } 1 \leq i \leq n$$

We call this interpretation of the activation states which is based on the ordering (2) the *Boolean option*.⁴

Sometimes it is useful to assume that both endpoints of the unit's working range carry maximal information and one value in the centre of the scale carries minimal information. The plausibility of such a choice was mentioned by Balkenius & Gärdenfors (1991). These authors suggested to take both $\mathbf{0}$ and $\mathbf{1}$ as states of maximal information and to assume that there is a resting state $\frac{1}{2}$ that represents minimal information. Unfortunately, they didn't work out this proposal.

In Blutner (2004) the working range of each unit is stipulated to be $[-1, +1]$; the activations $+1$ and -1 indicate maximal specification; the resting activation 0 indicates (complete) underspecification. Generalizing Balkenius & Gärdenfors' (1991) idea, the set $S = [-1, +1]^n$ of activation states can be partially ordered in accordance with their informational content:

$$(3) \quad s \geq t \text{ iff } s_i \geq t_i \geq 0 \text{ or } s_i \leq t_i \leq 0, \text{ for all } 1 \leq i \leq n. \quad (\text{Read } s \geq t \text{ as } s \text{ is at least as specific as } t)$$

It is a simple exercise to show that the poset $\langle S, \geq \rangle$ doesn't form a lattice yet. However, it can be extended to a lattice by introducing a set \perp of *impossible activation states*: $\perp = \{s: s_i = \text{nil} \text{ for } 1 \leq i \leq n\}$, where *nil* designates the "impossible" activation of an unit, i.e. a clash between positive and negative activation (for details, see Blutner, 2004). Further, it is possible to show that the extended poset of activation states $\langle S \cup \perp, \geq \rangle$ forms a DeMorgan lattice. This allows us to interpret these activation states as propositional objects ('information states'). It is convenient to call this interpretation of the activation states the *DeMorgan option*.⁵

Symmetric networks may be viewed as searching for the local minima of a quadratic function called an energy function (or Ljapunov function). The

⁴ $\langle S, \geq \rangle$ forms a Boolean algebra if the underlying neural network is binary (cf. Balkenius & Gärdenfors, 1991)

⁵ There is another option for modelling activation states: ortho-algebras. The interested reader could consult www.quantum-cognition.de in order to learn more about this alternative option. Unfortunately, space limitation forbids us to discuss the ortho-algebraic approach in the present article.

important fact proven in Hopfield (1982) says that in the case of asynchronous (non-deterministic) updates, the function

$$(4) \quad E(s) = -\sum_{i>j} w_{ij} s_i s_j$$

is a Ljapunov function of the dynamic system described by the equation in (1)⁶; i.e., when the activation state of the network changes, E can either decrease or remain the same. Hence, the output states $\lim_{n \rightarrow \infty} f^n(s)$ can be characterized as *the local minima* of the Ljapunov-function. A consequence of this result is that all states s in a symmetric network develop under asynchronous updating into *resonances*, i.e. into stable states of the network that *attract* other states (for details, see Cohen & Grossberg, 1983).

Usually, asynchronous updating results in stable states that are local but not global minima of the energy function E . The Boltzman machine (Hinton & Sejnowski, 1983; Hinton & Sejnowski, 1986) is a modification of the Hopfield network that realizes the *global* minima, i.e. their output states $\lim_{n \rightarrow \infty} f^n(s)$ can be characterized as *the global minima* of the Ljapunov-function. Like the Hopfield net, the Boltzman machine updates its units by means of an asynchronous update procedure. However, it employs a stochastic activation function rather than a deterministic one. This activation function can be considered to realize some stochastic noise ("faults") in a decreasing rate during the processing of a single pattern.⁷

Updating an information state s may result in an information state $f...f(s)$ that does not include the information of s . However, if we want to handle logical inferences, it is important to interpret updating as specification. That means we have to make sure that the initial state s has to be informationally included in the resulting update. Hence, we have to "clamp" s somehow in the network. A technical way to do that has been proposed by Balkenius & Gärdenfors (1991) making use of an update function \underline{f} that 'clamps' s in the network (see also Blutner, 2004).⁸ Fortunately, the aforementioned formal results derived for asymptotic updating without clamping also hold for asynchronous updating with clamping.

Hence, the following set of *asymptotic updates* of s is well defined if we use an asynchronous update function \underline{f} with clamping:

⁶ The simple form of the energy function is due to assuming zero thresholds. We can always mimic the case of non-zero thresholds by assuming bias nodes with a fixed input activation.

⁷ The procedure is called 'simulated annealing' (based on an analogy from physics). For details see Hinton & Sejnowski, (1983; 1986).

⁸ Clamping is not only required if we try to model logical inferences in a connectionist network but also applies in the case of pattern completion (see, e.g. Rumelhart, Hinton, & McClelland, 1986; Smolensky, 1986).

$$(5) \quad \text{ASUP}_w(s) = \{t: t = \lim_{n \rightarrow \infty} \underline{f}^n(s)\}$$

Further, in the case of the Boltzman machine, we can characterize the set of asymptotic updates as the set of all specifications of s that minimize the energy E of the system. Using the expression $\min_E(s)$ to indicate this set of global energy minima, we have

$$(6) \quad \text{ASUP}_w(s) = \min_E(s).$$

The following example (borrowed from Blutner, 2004) gives an illustration of the basic concepts introduced so far.

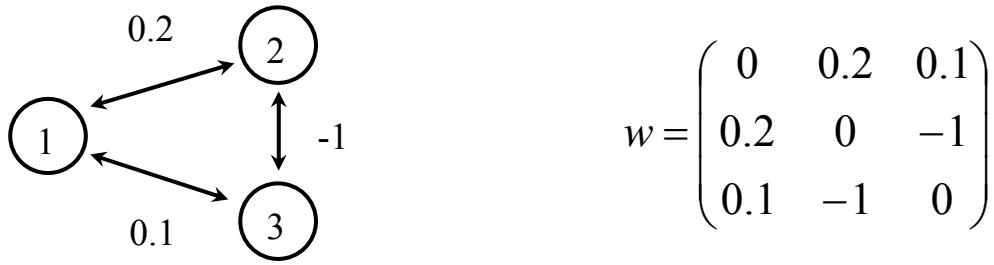


Figure 1: Symmetric network with weight matrix

This figure shows a symmetric network consisting of three units (labelled 1, 2, and 3) and the corresponding connection matrix w . The set of activation states is $S = [-1, +1]^3$. Clamping node 1, the fast dynamics yields an output state where node 2 is activated and node 3 is inhibited:

$$(7) \quad \text{ASUP}_w(<1 \ 0 \ 0>) = \{<1 \ 1 \ -1>\}$$

The same result is obtained if we consider the energy function on the domain S :

$$(8) \quad E(s) = -0.2 s_1 s_2 - 0.1 s_1 s_3 + s_2 s_3$$

The following table shows the nine possible specifications of the initial state $<1 \ 0 \ 0>$ if we restrict ourselves to the discrete subdomain $S' = \{-1, 0, 1\}^3$:

Table 1: Discrete specifications of $\langle 1 \ 0 \ 0 \rangle$ and the energy of all specifications. The energy-minimal state is indicated by \rightarrow . It corresponds to the output state given in (7).

s [state]	$E(s)$ [energy]
$\langle 1 \ 0 \ 0 \rangle$	0
$\langle 1 \ 0 \ 1 \rangle$	-0.1
$\langle 1 \ 0 \ -1 \rangle$	0.1
$\langle 1 \ 1 \ 0 \rangle$	-0.2
$\langle 1 \ 1 \ 1 \rangle$	0.7
$\langle 1 \ 1 \ -1 \rangle$	-1.1 \rightarrow
$\langle -1 \ -1 \ 0 \rangle$	0.2
$\langle -1 \ -1 \ 1 \rangle$	-0.9
$\langle -1 \ -1 \ -1 \rangle$	1.3

In order to demonstrate that the working range of the nodes of the network is not essential for the dynamic properties of the network, we modify our example so that it relates to an activation space $[0, 1]^3$. The discrete subspace that corresponds to the states in Table 1 is obtained if we consider the map $1 \rightarrow 1$, $0 \rightarrow \frac{1}{2}$, and $-1 \rightarrow 0$. Further, we have to adapt the energy function from (8) which is based on zero thresholds. Instead of the zero thresholds we assume thresholds $\theta_i = \frac{1}{2}$, which are positioned in the centre of the working range. As a consequence, we have to add an additional term $-\sum_i \theta_i \cdot s_i$, which can also be seen as a consequence of introducing bias nodes with input activity 1 (see footnote 5):

$$(9) \quad E(s) = -0.2 s_1 s_2 - 0.1 s_1 s_3 + s_2 s_3 - \frac{1}{2} (s_1 + s_2 + s_3)$$

Table 2 shows the energies of the corresponding states of the discrete subspace $\{0, \frac{1}{2}, 1\}^3$. As a matter of fact the energy ordering of the states in Table 2 is the same as the energy ordering of the corresponding states in Table 1. Hence, the working space of the neurons does not really affect the ordering of the states if the thresholds are adopted accordingly.

Table 2: Corresponding specifications for the activation space $[0, 1]^3$. The energy is calculated according to formula (9) and the energy of all specifications. The energy-minimal state is indicated by \rightarrow .

s [state]	$E(s)$ [energy]
$\langle 1 \frac{1}{2} \frac{1}{2} \rangle$	-0.9
$\langle 1 \frac{1}{2} 1 \rangle$	-0.95
$\langle 1 \frac{1}{2} 0 \rangle$	-0.85
$\langle 1 1 \frac{1}{2} \rangle$	-1.00
$\langle 1 1 1 \rangle$	-0.8
$\langle 1 1 0 \rangle$	-1.2 \rightarrow
$\langle 1 0 \frac{1}{2} \rangle$	-0.8
$\langle 1 0 1 \rangle$	-1.1
$\langle 1 0 0 \rangle$	-0.5

Although the actual working range of a unit is only of marginal interest, the interpretation of the activation values is essential. If we take the interval $[0, 1]$ as working range, for instance, then the interpretation of the value 0 is essential. We can either see 0 as indicating maximal underspecification or as indicating maximal specification (together with the value 1; the value $\frac{1}{2}$ is typically used to indicate underspecification in this case). The former interpretation conforms to the Boolean option; the latter conforms to the DeMorgan option. The consequences of this distinction are discussed in sections 4 and 5.

3 Examples

In the previous section we have seen that the propositional objects called information states are related by a partial ordering \geq . It is obvious that this relation can be interpreted as a strict (monotonic) entailment relation since it satisfies the Tarskian restrictions for such a relation:

- (10) a. $s \geq s$ (Reflexivity)
b. if $s \geq t$ and $s \circ t \geq u$, then $s \geq u$ (Cut)
c. if $s \geq u$, then $s \circ t \geq u$ (Monotonicity)

Here we have to make use of the operation $s \circ t =_{\text{def}} \sup\{s, t\}$, which is called *conjunction*. This operation expresses the *simultaneous realization* of two activation states. In the case where \geq expresses the positive informational

content with regard to the state set $[0, 1]^n$ (*Boolean option*) the explicit form of the conjunction operation is given in (11a); in the second case where \geq expresses specificity with regard to the state set $[-1, 1]^n$ (*DeMorgan option*) the conjunction operation is given in (11b):

$$(11) \quad \begin{array}{ll} \text{a.} & (s \circ t)_i = \max(s_i, t_i) \\ \text{b.} & (s \circ t)_i = \begin{cases} \max(s_i, t_i), & \text{if } s_i, t_i \geq 0 \\ \min(s_i, t_i), & \text{if } s_i, t_i \leq 0 \\ \text{nil}, & \text{elsewhere} \end{cases} \end{array}$$

As shown by Balkenius & Gärdenfors (1991), Blutner (2004), and in a somewhat different sense by Hölldobler (1991), Pinkas (1995), and others, it is possible to define a nonmonotonic inference relation that reflects asymptotic updating of information states. Let $\langle S, \geq \rangle$ be a poset of activation states, and w the connection matrix. Then the notion of asymptotic updates as given in (5) naturally leads to a nonmonotonic inferential relation between information states defined as follows (cf. Blutner, 2004):

$$(12) \quad s \approx_w t \text{ iff } s' \geq t \text{ for each } s' \in \text{ASUP}_w(s)$$

Of course, there is an equivalent formulation in terms of energy minimization:⁹

$$(13) \quad s \approx_E t \text{ iff } s' \geq t \text{ for each } s' \in \min_E(s)$$

We also call the inferential relation between information states *subsymbolic inferential relation* and the inferences themselves *subsymbolic inferences*.

Following Balkenius & Gärdenfors (1991), the inferential notion that is adequate to describe how neuron activities spread through the network (i.e. the *fast dynamics* of a neural system) can be characterized in terms of the general postulates that Gabbay (1985) and Kraus, Lehmann, and Magidor (1990) have seen as constituting a *cumulative* (nonmonotonic) consequence relation. This holds independently of the particular working range that is chosen for the nodes of the network and it rests on the equivalence of the two inferential notions defined in (12) and (13). In (14) the relevant properties are listed.

$$(14) \quad \begin{array}{ll} \text{a.} & \text{if } s \geq t, \text{ then } s \approx_w t & (\textit{Supraclassicality}) \\ \text{b.} & s \approx_w s & (\textit{Reflexivity}) \end{array}$$

⁹ We simply have to use of the equivalence (6) that holds in the case of the Boltzman machine.

- c. if $s \approx_w t$ and $s \circ t \approx_w u$, then $s \approx_w u$ (*Cut*)
- d. if $s \approx_w t$ and $s \approx_w u$, then $s \circ t \approx_w u$ (*Cautious Monotonicity*)

For a proof of the validity of these properties in the case of a symmetric network, see Blutner (2004).

Going back to the earlier example introduced in Figure 1, it is a simple exercise to show that the following inferences are valid:

- (15) a. $\langle 1 \ 0 \ 0 \rangle \approx_w \langle 1 \ 1 \ -1 \rangle$
 b. $\langle 1 \ 0 \ 0 \rangle \approx_w \langle 1 \ 1 \ 0 \rangle$
 c. $\langle 1 \ 0 \ 0 \rangle \approx_w \langle 0 \ 1 \ 0 \rangle$

The latter two inferences can be derived from the first one by taking into account that $\langle 1 \ 1 \ -1 \rangle \geq \langle 1 \ 1 \ 0 \rangle \geq \langle 0 \ 1 \ 0 \rangle$.

In connectionist systems (domain) knowledge is encoded in the connection matrix w (or, alternatively, the energy function E). In the following two sections I want to discuss the close correspondence to certain symbolic systems that represent knowledge in a database consisting of expressions with default status.

4 Penalty Logic

According to Pinkas (1992, 1995), domain knowledge can be represented by a logic-based scheme, the *Penalty Logic*. This logic associates to each formula of a knowledge base the price to pay if this formula is violated. In this section I will give a concise introduction into Penalty Logic following in part the exposition in de Saint-Cyr, Lang, & Schiex (1994). Further, I will make clear that we have to adopt the Boolean option of interpreting activation states in order to reconstruct Pinkas' claim of the equivalence between inferences in Penalty Logic and inferences in symmetric networks.

Let's consider the language \mathcal{L}_{At} of propositional logic (referring to the alphabet At of atomic symbols). A triple $\langle At, \Delta, k \rangle$ is called a *penalty knowledge base* (PK) iff (i) Δ is a set of consistent sentences built on the basis of At (the possible hypotheses); (ii) $k: \Delta \Rightarrow (0, \infty)^{10}$ (the penalty function). Intuitively, the penalty of an expression δ represents what we should pay in order to get rid of δ . If we pay the requested price we no longer have to satisfy δ . Hence, the larger $k(\delta)$ is, the more important δ is.

Let α be a formula of our propositional language \mathcal{L}_{At} . A *scenario*¹¹ of α in PK is a subset Δ' of Δ such that $\Delta' \cup \{\alpha\}$ is consistent. The cost $K_{PK}(\Delta')$ of a

¹⁰ The notation $(0, \infty)$ refers to the positive real numbers (excluding 0).

¹¹ I borrow this expression from Poole (1988).

scenario Δ' in PK is the sum of the penalties of the formulas of PK that are not in Δ' :

$$(16) \quad K_{PK}(\Delta') = \sum_{\delta \in (\Delta - \Delta')} k(\delta)$$

A *optimal scenario of α in PK* is a scenario the cost of which is not exceeded by any other scenario (of α in PK), so it is a penalty minimizing scenario. With regard to a penalty knowledge base PK, the following cumulative consequence relation can be defined:

$$(17) \quad \models_{PK} \beta \text{ iff } \beta \text{ is an ordinary consequence of each optimal scenario of } \alpha \text{ in PK.}$$

Hence, penalties may be used as a criterion for selecting preferred consistent subsets in an inconsistent knowledge base, thus inducing a non-monotonic inference relation.

To illustrate the approach I consider an example from Asimov (1950). Isaac Asimov described what became the most famous view of the ethical rules for robot behaviour in his “three laws of robotics”¹²:

First Law

A robot may not injure a human being.¹³

Second Law

A robot must follow (obey) the orders given it by human beings, except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Now assume some human X says to the robot 'kill my wife'. The relevant knowledge base can be formalized by five propositional formulae, where I, F, P have the obvious intended meaning in connection with the three laws, S expresses that some human X gives this shocking order to the robot, and K expresses the content of the order. The first three formulae in (18) express the three laws, the last two formulae express very strong meaning postulates:

¹² Thanks to Bart Geurts for drawing my attention to this example.

¹³ I am simplifying a bit. The original clause is more complicated: "A robot may not injure a human being, or, through inaction, allow a human being to come to harm."

(18)	$\neg I$	5
	F	2
	P	1
	$(S \wedge F) \rightarrow K$	1000
	$K \rightarrow I$	1000

The positive real numbers associated with the formulae are the penalties. Consider now the following two scenarios for S:

$$(19) \quad \Delta_1 = \{\neg I, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$$

$$\Delta_2 = \{F, P, (S \wedge F) \rightarrow K, K \rightarrow I\}$$

The cost of these two scenarios with regard to the PK given in (19) are $K_{PK}(\Delta_1) = 2$ and $K_{PK}(\Delta_2) = 5$, respectively. Since the cost of all other possible scenarios is higher, we can conclude that Δ_1 is the optimal scenario of S. Hence, according to the ethical rules, our robot should not injure anybody, neither X's wife nor himself.

Now we come to the semantic interpretation of the Penalty Logic introduced so far. Let v denote an ordinary (total) interpretation for the language \mathcal{L}_{At} ($v: At \rightarrow \{0,1\}$). The usual clauses apply for the evaluation $\llbracket \cdot \rrbracket_v$ of the formulas of \mathcal{L}_{At} relative to v . The following function indicates how strongly an interpretation v conflicts with the space of hypotheses Δ of a penalty knowledge base PK:

$$(20) \quad \mathcal{E}_{PK}(v) =_{\text{def}} \sum_{\delta \in \Delta} k(\delta) \llbracket \neg \delta \rrbracket_v \quad (\mathcal{E} \text{ is called the } \textit{system energy} \text{ of the interpretation})^{14}$$

An interpretation v is called a *model* of α just in case $\llbracket \alpha \rrbracket_v = 1$. A *preferred model* of α is a model of α with minimal energy \mathcal{E} (with regard to the other models of α). As the semantic counterpart to the syntactic notion $\alpha \sim_{PK} \beta$ given in (17) we can define the following relation:

$$(21) \quad \alpha \approx_{PK} \beta \text{ iff each preferred model of } \alpha \text{ is a model of } \beta.$$

As a matter of fact, the syntactic notion (17) and the semantic notion (21) coincide. Hence, the logic is sound and complete. A proof can be found in Pinkas (1995).

¹⁴ What I call the system energy of an interpretation (with regard to a PK) is called *violation rank* for the interpretation in Pinkas (1995); deSaint-Cyr et al. (1994) call it the *cost of interpretation*.

With regard to the integration of neural networks and symbolic systems, Pinkas (1992, 1995) made a breakthrough. On the one hand he was able to demonstrate that the problem of finding preferred models for a given set of assumptions can be reduced to the minimization problem of an energy function in symmetric networks. On the other hand he showed that the minimization problem of an energy function of a symmetric network can be reduced to the problem of finding preferred models for a given set of assumptions representing domain knowledge

In the following I will give a concise description of Pinkas' basic results. I start with sketching the transformation that enables one to construct a symmetric network that is *strongly equivalent* with a given knowledge base PK. Strong equivalence means that the energy function of the neural network and the system energy of the knowledge base in Penalty Logic are the same (up to a constant c). I will sketch the basic elements of this transformation only; the reader is referred to Pinkas (1992; 1995) for a fuller description.

For each logical expression α a characteristic function $B(\alpha): [0, 1]^n \rightarrow [0, 1]$ is defined. The letter B for the translation operation indicates that the translation relates to the *Boolean option* of interpreting activation states. The characteristic function $B(\alpha)$ is defined in its analytical form making use of variables x_i which refer to real numbers in the interval $[0, 1]$.

- (22) a. $B(p_i) = x_i$, where p_i designates the i^{th} atomic formula of \mathcal{L}_{At}
- b. $B(\neg\alpha) = 1 - B(\alpha)$
- c. $B(\alpha \wedge \beta) = B(\alpha) \cdot B(\beta)$

It is simple to see the characteristic function $B(\alpha)$ has its maximum value(s) exactly when α has a value of true (supposing the integer values of x_i are the values of the interpretations of p_i). For example, $B(p_1 \wedge p_2) = x_1 \cdot x_2$.¹⁵ The maximization of $x_1 \cdot x_2$ yields $x_1 \rightarrow 1, x_2 \rightarrow 1$. Further, $B(p_1 \rightarrow p_2) = B(\neg(p_1 \wedge \neg p_2)) = x_1 \cdot x_2 - x_1 + 1$ and the maximization of the resulting term gives three solutions corresponding to the three interpretations that make the material implication true. Finally, $B(p_1 \vee p_2) = B(\neg(\neg p_1 \wedge \neg p_2)) = x_1 \cdot x_2 - x_1 - x_2$; the maximization again gives three solutions. Figure 2 provides a graphical representation of the three characteristic functions.

¹⁵ The same function is sometimes used in fuzzy logic. It is called product t-norm (cf. Hajek, 1998).

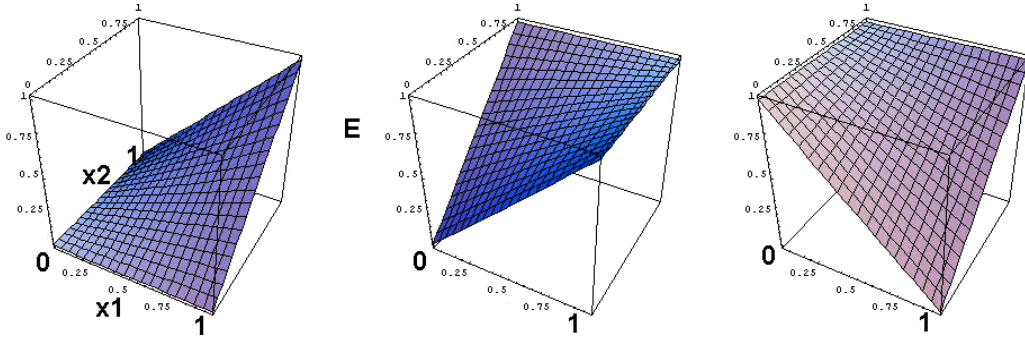


Figure 2: graphical representation of the characteristic functions for conjunction, disjunction, and material implication, respectively (from left to right)

Now we are ready to introduce a translation of a *penalty knowledge base* $\langle \Delta, \Delta, k \rangle$ into a symmetric network. We simply construct a network with the following energy function using the characteristic function B for translating propositional formulas into numerical functions:

$$(23) \quad E(x_1, \dots, x_n) = \sum_{\delta \in \Delta} k(\delta) \cdot B(\neg \delta)$$

It can be shown that the constructed symmetric network is *strongly equivalent* with the given knowledge base PK. In other words, we have the following fact:

Fact 1:

For each knowledge base PK with the assigned energy function E:

$\mathcal{E}_{PK}(v) = E(x_1, \dots, x_n)$ for each interpretation v provided $v(p_i) = x_i$

The proof is a simple consequence of the observation that the value of a propositional formula δ for a given interpretation v is the same as the value of the corresponding characteristic function $B(\delta)$ provided $v(p_i) = x_i$, i.e.

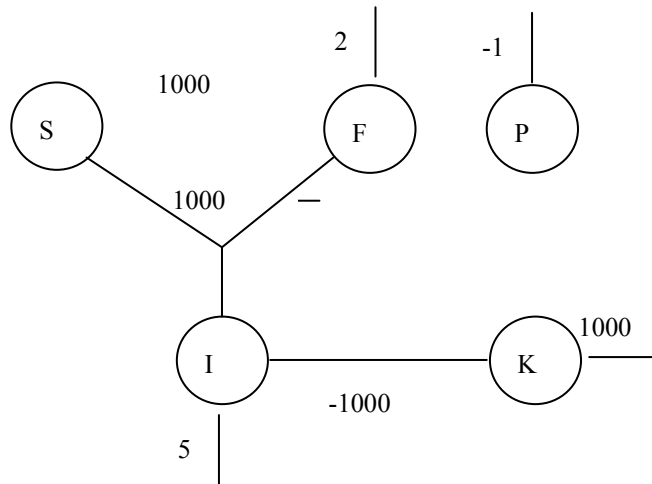
$$(24) \quad \llbracket \delta \rrbracket_v = B(\delta) [v(p_1)/x_1, \dots, v(p_n)/x_n]$$

Fact 1 then immediately follows from the definition of \mathcal{E} given in (20). The proof of (24) is by induction using the translation provided in (22). Taking up the earlier example about the robot's ethics (18), we come to the following energy calculation (instead of the variables x_i we use the names of the atomic formulas as names for the variables):

Table 3: Calculation of the energy function for the PK given in (19)

Penalty	Expression in PK	Energy function
5	$\neg I$	$5 I$
2	F	$-2F$
1	P	$-P$
1000	$(S \wedge F) \rightarrow K$	$1000(S \cdot F - S \cdot F \cdot I)$
1000	$K \rightarrow I$	$1000(K - K \cdot I)$
		$E = 5I - 2F - P + 1000K + 1000S \cdot F - 1000K \cdot I - 1000S \cdot F \cdot I$

The energy function contains a cubic term $-1000S \cdot F \cdot I$ that goes beyond the simple quadratic energy functions introduced in (4). Such higher order energy functions refer to connectionist networks having sigma-pi units with multiplicative connections (Rumelhart et al., 1986). In the case under discussion, the following network results:


Figure 3: Higher order network representing the energy function calculated in Table 3

Pinkas (1992) has shown that higher order terms can be eliminated by introducing *hidden units*. In the case of the cubic terms $\text{const} \cdot X \cdot Y \cdot Z$ here is the relevant elimination rule, where the variable T refers to the hidden unit:

$$(25) \quad w \cdot X \cdot Y \cdot Z = \begin{cases} 2w \cdot X \cdot T + 2w \cdot Y \cdot T + 2w \cdot Z \cdot T - 5w \cdot T, & \text{if } w < 0 \\ w \cdot X \cdot Y - 2w \cdot X \cdot T - 2w \cdot Y \cdot T + 2w \cdot Z \cdot T + 3w \cdot T, & \text{if } w > 0 \end{cases}$$

In the present case the coefficient is negative and the final quadratic energy function is

$$(26) E = 5I - 2F - P + 1000SF - 2000ST - 2000FT - 2000IT + 5000T + 1000K - 1000KI$$

The final network with quadratic the energy function and the hidden node T is shown in Figure 4.

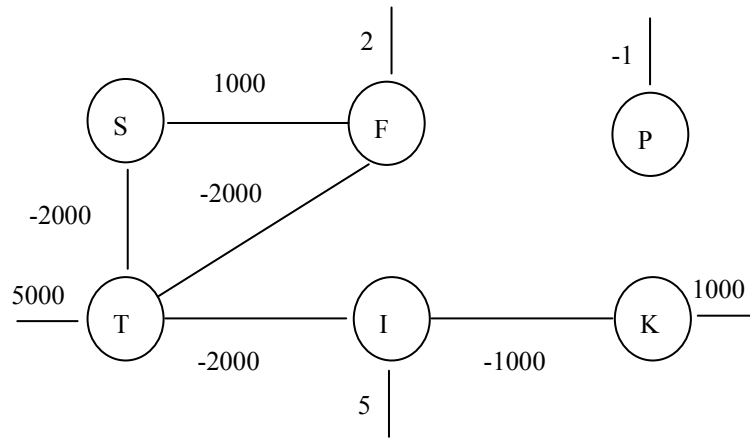


Figure 4: First order network with one hidden unit T

This was my brief sketch of how to translate any knowledge base PK into a strongly equivalent symmetric network supposed the Boolean option of interpreting activation states has been adopted.

There is also a reverse procedure that translates any symmetric network into a PK. I will outline this translation now. For simplicity, I exclude higher order units and/or hidden units. We consider a Hopfield system with connection matrix w (n units), and we assume $At = \{p_1, \dots, p_n\}$ to be a set of atomic symbols. Then we consider the following formulae β_{ij} of \mathcal{L}_{At} :

$$(27) \quad \beta_{ij} =_{\text{def}} \text{sign}(w_{ij})(p_i \wedge p_j), \text{ for } 1 \leq i < j \leq n \quad ^{16}$$

For each connection matrix w the *associated penalty knowledge base* is defined as $PK_w = \langle At, \Delta_w, k_w \rangle$, where the following two clauses apply:

¹⁶ $\text{Sign}(x)$ is an operator that introduces a negation sign " \neg " for $x < 0$ and it leaves the expression in its scope unchanged if $x \geq 0$. For instance, $\text{sign}(0.2)(\alpha) = \alpha$ and $\text{sign}(-0.2)(\alpha) = \neg\alpha$.

$$(28) \quad \begin{array}{ll} \text{a.} & \Delta_w = \{ \beta_{ij} : 1 \leq i < j \leq n \} \\ \text{b.} & k_w(\beta_{ij}) = |w_{ij}| \end{array}$$

With these notations at hand we can state the following fact:

Fact 2:

For each connection matrix w the energy function $E(s) = -\sum_{i>j} w_{ij} s_i s_j$ is strongly equivalent with the associated penalty knowledge base PK_w ; i.e. $\mathcal{E}_{PK}(v) = E(s_1, \dots, s_n) + \text{constant}$, provided $v(p_i) = s_i$

For the proof we notice first that

$$[\![\beta_{ij}]\!]_v = [\![\text{sign}(w_{ij})(p_i \wedge p_j)]\!]_v = \begin{cases} v(p_i) \cdot v(p_j), & \text{if } w_{ij} \geq 0 \\ 1 - v(p_i) \cdot v(p_j), & \text{if } w_{ij} < 0 \end{cases}$$

Then we have the following equivalences: $\mathcal{E}_{PK}(v) =_{\text{def}} \sum_{\delta \in \Delta} k(\delta) [\![\neg \delta]\!]_v = \sum_{i>j} k(\beta_{ij}) [\![\neg \beta_{ij}]\!]_v = \text{const} - \sum_{i>j} w_{ij} \cdot v(p_i) \cdot v(p_j) = \text{const} + E(s) + \text{constant}$ (provided $v(p_i) = s_i$). Hence, $\mathcal{E}_{PK}(v)$ and $E(s)$ differ only by a term $\text{const} = \frac{1}{2} \sum_{i>j} (w_{ij} + |w_{ij}|)$ and are therefore strongly equivalent.

For the example introduced in Figure 1 the energy function (9) was associated assuming bias nodes with fixed activity 1 that mimic thresholds $\theta_i = \frac{1}{2}$. This expression is repeated here for convenience:

$$(9) \quad E(s) = -0.2 s_1 s_2 - 0.1 s_1 s_3 + s_2 s_3 - 0.5 (s_1 + s_2 + s_3)$$

The associated penalty knowledge base then comes out as follows:

$$(29) \quad \begin{array}{ll} p_1 \wedge p_2 & 0.2 \\ p_1 \wedge p_3 & 0.1 \\ \neg(p_2 \wedge p_3) & 1 \\ p_1 & 0.5 \\ p_2 & 0.5 \\ p_3 & 0.5 \end{array}$$

With regard to this PK it is not difficult to show that

$$(30) \quad \begin{array}{ll} \text{a.} & p_1 \mid \sim_{PK} p_2 \\ \text{b.} & p_1 \mid \sim_{PK} \neg p_3 \end{array}$$

It would be nice to have a possibility to express such inferences directly as subsymbolic inferences in the corresponding network. Unfortunately, this is possible only for inferences between positive literals such as considered in (30a):

$$(31) \quad \langle 1 \ 0 \ 0 \rangle \approx_E \langle 1 \ 1 \ 0 \rangle$$

Here the state $\langle 1 \ 0 \ 0 \rangle$ indicates an activation of the first node that corresponds to the atom p_1 , and $\langle 1 \ 1 \ 0 \rangle$ indicates that, in addition, the second node is activated (corresponding to p_2). Unfortunately, the zero elements cannot be interpreted as negations. The reason is that in the Boolean option of interpreting node activities the vector $\mathbf{0} = \langle 0, 0, \dots, 0 \rangle$ indicates a scheme with minimal informational content. Hence, 0 indicates maximum underspecification, not a negative truth-value. As a consequence, we have no direct way to express the inferences (30b) in the subsymbolic mode.¹⁷ In the next section we overcome this shortcoming by adopting the DeMorgan option of interpreting activation states.

5 Penalty/Reward Logic

The DeMorgan option of interpreting activation states means that we explicitly consider a resting state in the *centre* of the unit's working range in order to represent minimal information (complete underspecification). For reasons of symmetry and parsimony I choose the interval $[-1, +1]$ as working range of a unit; the activations $+1$ and -1 indicate maximal specification (corresponding to the truth values T and F); the activation 0 indicates underspecification (see Section 2).

Assuming a symmetric network with n nodes it is possible now to express *all* elements of the discrete subspace $\{-1, 0, +1\}^n \subset [-1, 0, +1]^n$ by symbolic expressions. Following Blutner (2004), we can do this formally by interpreting the conjunction of literals in \mathcal{L}_{At} by the corresponding elements of the DeMorgan algebra $\langle S \cup \perp, \geq \rangle$. More precisely, we call the triple $\langle S \cup \perp, \geq, \uparrow \rangle$ a *Hopfield model* for \mathcal{L}_{At} if and only if $\uparrow \downarrow$ is a function assigning some element of $S \cup \perp$ to each atomic symbol and obtaining the following conditions:

¹⁷ Of course, we can introduce a hard rule $\neg p_3 \leftrightarrow p_4$ in the knowledge base PK, and correspondingly an additional node that corresponds to p_4 into the network. Then we have $p_1 \vdash_{PK} p_4$ instead of (30b) and this corresponds to $\langle 1 \ 0 \ 0 \ 0 \rangle \approx_E \langle 0 \ 0 \ 0 \ 1 \rangle$ in the extended space.

- (32) a. $\models \alpha \wedge \beta \models = \models \alpha \wedge \models \beta \models$
 b. $\models \neg \beta \models = -\models \beta \models$ (" \neg " converts positive into negative activation and *vice versa*).

A Hopfield model is called *local* (for \mathcal{L}_{At}) iff it realizes the following assignments:

- (33) $\models p_1 \models = \langle 1 \ 0 \ \dots \ 0 \rangle$
 $\models p_2 \models = \langle 0 \ 1 \ \dots \ 0 \rangle$
 \dots
 $\models p_n \models = \langle 0 \ 0 \ \dots \ 1 \rangle$

An information state s is said to be *represented* by a formula α of \mathcal{L}_{At} (relative to a Hopfield model M) iff $\models \alpha \models = s$. It is obvious that each discrete state $s \in \{-1, 0, +1\}^n$ can be represented by a conjunction of literals in \mathcal{L}_{At} using the local Hopfield model M given in (33). For instance, if we take $n=3$, the following formulae *represent* proper activation states: (i) p_1 represents $\langle 1 \ 0 \ 0 \rangle$, (ii) p_2 represents $\langle 0 \ 1 \ 0 \rangle$, (iii) p_3 represents $\langle 0 \ 0 \ 1 \rangle$, (iv) $p_1 \wedge p_2$ represents $\langle 1 \ 1 \ 0 \rangle$, (v) $\neg p_1$ represents $\langle -1 \ 0 \ 0 \rangle$, and (vi) $p_1 \wedge p_2 \wedge \neg p_3$ represents $\langle 1 \ 1 \ -1 \rangle$. Hence, for local Hopfield models each discrete activation state can be considered symbolic.

Now the following important question arises: can each connection matrix be translated into domain knowledge such that all subsymbolic inferences between information states correspond to inferences in a certain symbolic system (perhaps a Penalty Logic or a modification of it)? And, conversely: can we translate domain knowledge into a connection matrix such that all symbolic inferences of our logical system correspond to subsymbolic inferences of the connectionist system? The answer to both these questions is *yes* if we use a variant of Pinkas' Penalty Logic – a variant I will call *Penalty/Reward Logic*. I will proceed as follows: first I introduce Penalty/Reward Logic, next I explain the transformation that encodes domain knowledge expressed in this logical system into a connection matrix of a symbolic network, after that I present the reverse transformation, and finally I discuss the advantages of the present approach in comparison with Pinkas' approach.

The syntax of Penalty/Reward Logic is the same as the syntax of Penalty Logic. Hence, we consider the language \mathcal{L}_{At} of propositional logic (referring to the alphabet At of atomic symbols) and take a triple $\langle At, \Delta, k \rangle$ as a *penalty/reward knowledge base* (PRK) where (i) Δ is a set of consistent sentences built on the basis of At and (ii) $k: \Delta \Rightarrow (0, \infty)$ is our cost function. The

idea that is connected with the cost function is that it penalizes an expression of Δ if it is not satisfied with regard to given circumstances and it rewards an expression of Δ if it is satisfied. Hence, for a *scenario of α in PRK* (i.e. a subset Δ' of Δ such that $\Delta' \cup \{\alpha\}$ is consistent) the cost $K_{PRK}(\Delta')$ of the scenario Δ' is defined as follows:

$$(34) \quad K_{PRK}(\Delta') =_{\text{def}} \sum_{\delta \in (\Delta - \Delta')} k(\delta) - \sum_{\delta \in \Delta'} k(\delta)$$

Hence, the cost of a scenario takes into account both the beliefs that are included in the scenario Δ' and the beliefs that are not included in Δ' . The missing beliefs give a positive contribution to the overall cost and the included beliefs give a negative contribution. This contrasts with the Penalty Logic correspondence (16) where only the missing beliefs count.

However, this contrast is not really striking since we can show that Penalty Logic and Penalty/Reward Logic are weakly equivalent in the terminology of Pinkas (1995); that means they are connected by a linear transformation:

$$(35) \quad K_{PRK}(\Delta') = 2 K_{PK}(\Delta') - \sum_{\delta \in \Delta} k(\delta)$$

The last term can be seen as constant. As a consequence, Penalty Logic and Penalty/Reward Logic produce the same orderings of scenarios. However, there are differences in the probability distributions that can be calculated by using standard statistical techniques (Boltzman machine: cf. Hinton & Sejnowski, 1983; Hinton & Sejnowski, 1986).

I will define now the system energy $\mathcal{E}_{PRK}(v)$ which indicates how strongly an interpretation v conflicts with the space of hypotheses Δ of the knowledge base PRK:

$$(36) \quad \mathcal{E}_{PRK}(v) =_{\text{def}} -\sum_{\delta \in \Delta} k(\delta) \llbracket \delta \rrbracket_v$$

This definition appears to be identical with the earlier definition (20). However, we are working with the DeMorgan option now and an interpretation v according to this option denotes a function $v: At \rightarrow \{-1, 1\}$. The usual clauses apply for the evaluation $\llbracket . \rrbracket_v$ of the formulas of \mathcal{L}_{At} relative to v if we take into account that -1 stands for *false* now instead of 0 in the Boolean case.

The definition (17) for a syntactic consequence relation and (21) for its semantic pendant can be taken over from the Boolean to the DeMorgan option:

$$(37) \quad \vdash_{PRK} \beta \text{ iff } \beta \text{ is an ordinary consequence of each optimal scenario of } \alpha \text{ in PRK (minimizing the cost } K_{PRK})$$

- (38) $\alpha \approx_{\text{PRK}} \beta$ iff each preferred model of α (minimizing the system energy \mathcal{E}_{PRK}) is a model of β .

As in the former case, the syntactic notion (37) and the semantic notion (38) coincide. Hence, the logic is sound and complete. A proof can be found in Blutner (2004).

Now I come to the transformation that enables one to construct a symmetric network that is *strongly equivalent* with a given knowledge base. Given a logical expression α a characteristic function $M(\alpha): [-1, 1]^n \rightarrow [-1, 1]$ is defined. The letter M indicates that the translation relates to the *DeMorgan option* of interpreting activation states. In the present case the generated variables x_i refer to real numbers in the interval $[-1, 1]$.

- (39) a. $M(p_i) = x_i$, where p_i designates the i^{th} atomic formula of \mathcal{L}_{At}
 b. $M(\neg\alpha) = -M(\alpha)$
 c. $M(\alpha \wedge \beta) = \frac{1}{2} (M(\alpha) \cdot M(\beta) + M(\alpha) + M(\beta) - 1)$

As a matter of fact the amount of the characteristic function $M(\alpha)$ has its *maximum* value exactly when α has a value of true (supposing the integer values of x_i are the values of the interpretations of p_i). For example, $M(p_1 \wedge p_2) = \frac{1}{2} (x_1 \cdot x_2 + x_1 + x_2 - 1)$. The maximization of $x_1 \cdot x_2 + x_1 + x_2 - 1$ yields $x_1 \rightarrow 1, x_2 \rightarrow 1$. Further, $M(p_1 \rightarrow p_2) = M(\neg(p_1 \wedge \neg p_2)) = \frac{1}{2} (x_1 \cdot x_2 + x_2 - x_1 + 1)$ and the maximization of the resulting term gives three solutions corresponding to the three interpretations that make the material implication true. For the disjunction we get $M(p_1 \vee p_2) = M(\neg(\neg p_1 \wedge \neg p_2)) = \frac{1}{2} (x_1 + x_2 - x_1 \cdot x_2 + 1)$; the maximization again gives three solutions. The shape of these functions is precisely as in Figure 2 but with axis values running from -1 to $+1$ instead of from 0 to 1 . It is further obvious that the characteristic function $M(\alpha)$ has its *minimum* value(s) exactly when α has a value of false. Now $\mathbf{0} = \langle 0 \ 0 \ 0 \rangle$ builds the centre of the three dimensional cube and it conforms to the point of maximum underspecification.

The translation that transforms a *penalty/reward knowledge base* $\langle \text{At}, \Delta, k \rangle$ into a symmetric network is straightforward. We simply construct a network with the following energy function using the characteristic function M for translating propositional formulas into numerical functions:

$$(40) \quad E(x_1, \dots, x_n) = -\sum_{\delta \in \Delta} k(\delta) \cdot M(\delta)$$

It can be shown that the constructed symmetric network is *strongly equivalent* with the given knowledge base PK. In other words, we have the following fact:

Fact 3:

For each knowledge base PRK with the assigned energy function E :

$\mathcal{E}_{\text{PRK}}(v) = E(x_1, \dots, x_n)$ for each interpretation v provided $v(p_i) = x_i$

As in the Boolean case, the proof is a consequence of the observation that the value of a propositional formula δ for a given interpretation v is the same as the value of the corresponding characteristic function $M(\delta)$ provided $v(p_i) = x_i$, i.e.

$$(41) \quad \llbracket \delta \rrbracket_v = M(\delta) [v(p_1)/x_1, \dots, v(p_n)/x_n]$$

Fact 3 then immediately follows from the definition of \mathcal{E}_{PRK} given in (36). The constructed network can contain higher order units. These units can be eliminated in the same way as discussed in section 4 by introducing hidden units. The main advantage of the DeMorgan option relates to the procedure that translates a symmetric network into a symbolic knowledge PRK. As in the Boolean case discussed before, I exclude higher order units and/or hidden units.

A connection between two nodes i and j contributes a term $w_{ij} \cdot x_i \cdot x_j$ to the energy function. Now we can ask what expression α translates to the product $x_i \cdot x_j$. The answer is the biconditional: $M(p_i \leftrightarrow p_j) = M((p_i \rightarrow p_j) \wedge (p_j \rightarrow p_i)) = x_i \cdot x_j + 1/8(x_i^2 \cdot x_j^2 - x_i^2 - x_j^2 + 1)$. The last term $1/8(x_i^2 \cdot x_j^2 - x_i^2 - x_j^2 + 1)$ can be neglected since it always gives the constant $1/8$ for the discrete values $\{-1, 0, 1\}$. Hence, I propose to consider the following expressions γ_{ij} as a translation of a single connection:

$$(42) \quad \gamma_{ij} =_{\text{def}} \text{sign}(w_{ij})(p_i \leftrightarrow p_j), \text{ for } 1 \leq i < j \leq n$$

For each connection matrix w the *associated penalty/reward knowledge base* is defined as $\text{PRK}_w = \langle \text{At}, \Delta_w, k_w \rangle$, where the following two clauses apply:

$$(43) \quad \begin{array}{ll} \text{a.} & \Delta_w = \{ \gamma_{ij} : 1 \leq i < j \leq n \} \\ \text{b.} & k_w(\gamma_{ij}) = |w_{ij}| \end{array}$$

Corresponding to fact 2 in the Boolean case, we can prove now the following fact (cf. Blutner 2004):

Fact 4:

For each connection matrix w the every energy function $E(s) = -\sum_{i > j} w_{ij} s_i s_j$ is strongly equivalent with the associated knowledge base PRK_w , i.e.

$\mathcal{E}_{\text{PK}}(v) = E(s_1, \dots, s_n) + \text{constant}$, provided $v(p_i) = s_i$

For the proof we notice first that $\llbracket \gamma_{ij} \rrbracket_v = \llbracket \text{sign}(w_{ij}) (p_i \leftrightarrow p_j) \rrbracket_v = \text{Sign}(w_{ij}) \cdot v(p_i) \cdot v(p_j)$, where $\text{Sign}(x)$ equals x if $x \geq 0$ and equals $-x$ if $x < 0$. Then we have the following equivalences: $\mathcal{E}_{\text{PRK}}(v) =_{\text{def}} -\sum_{\delta \in \Delta} k(\delta) \llbracket \delta \rrbracket_v = -\sum_{i>j} k(\gamma_{ij}) \llbracket \gamma_{ij} \rrbracket_v = -\sum_{i>j} |w_{ij}| \cdot \text{Sign}(w_{ij}) \cdot v(p_i) \cdot v(p_j) = -\sum_{i>j} w_{ij} \cdot v(p_i) \cdot v(p_j) = E(s)$. Hence, $\mathcal{E}_{\text{PRK}}(v)$ and $E(s)$ are identical provided $v(p_i) = s_i$. Thus, they are strongly equivalent.

At the beginning of this section we introduced local Hopfield models that allow one to represent each discrete information state by a conjunction of literals of the propositional language \mathcal{L}_{At} . Now we can state that each subsymbolic inference between information states corresponds to an inference in Penalty/Reward Logic (and vice versa). This is an immediate consequence of Facts 3 and 4.

Fact 5:

Let α and β be formulas that are conjunctions of literals. Assume further that a penalty/reward knowledge base PRK is associated with the connection matrix w – by using either the transformation $\text{PRK} \rightarrow w$ (40) or the transformation $w \rightarrow \text{PRK}$ (43). Then we have: $\models \alpha \models \models_w \models \beta$ iff $\alpha \models_{\text{PRK}} \beta$ (iff $\alpha \models_{\neg \text{PRK}} \beta$)

The equivalence between subsymbolic inferences in Hopfield networks and symbolic inferences in Penalty/Reward Logic can be applied in two different ways. First, this outcome of the integrative methodology can help the symbolist to find more efficient implementations of solving optimization problems and constraint satisfaction problems. Second, the results can help the connectionist to better understand their networks and to solve the so-called *extraction problem*, i.e the extraction of symbolic knowledge from connectionist networks. The latter approach was stressed by d'Avila Garcez, Broda, & Gabbay (2001) *inter alia*, the former was pioneered by Pinkas (1992, 1995).

In our example from Figure 1 the energy function (8) was calculated in case of the DeMorgan option, repeated here.

$$(8) \quad E(s) = -0.2 s_1 s_2 - 0.1 s_1 s_3 + s_2 s_3$$

The corresponding knowledge base is given by the following weight-annotated defaults.

$$(44) \quad \begin{array}{ll} p_1 \leftrightarrow p_2 & 0.2 \\ p_1 \leftrightarrow p_3 & 0.1 \\ p_2 \leftrightarrow \neg p_3 & 1 \end{array}$$

The translation mechanism is very simple and transparent: it translates a node i into the atomic symbol p_i , translates an activating link in the network into the logical biconditional \leftrightarrow , and translates an inhibitory link into the biconditional \leftrightarrow plus an internal negation \neg of one of its arguments. Furthermore, the weights of the defaults have to be taken as the absolute value of the corresponding matrix elements.

Is the difference between choosing the Boolean option and choosing the DeMorgan option really essential? A first hint for an essential difference is obtained if we look at Figure 5 which presents the energy function (8) as function of s_2 and s_3 with a fixed value $s_1=1$, i.e. the first node is clamped with its maximum activity. We are interested in calculating the minimum value of the energy regarding the s_2 - s_3 plane. Of course, the starting point for the minimization route is important. The De Morgan option allows us to take the starting point as expressing maximum underspecification. This corresponds to the vector $\langle 1 \ 0 \ 0 \rangle$ in the full three dimensional activation space or to the two dimensional projection $\langle 0 \ 0 \rangle$. This point is called B in Figure 5. B contrasts with the point A, which is $\langle -1 \ -1 \rangle$. A is the starting point in a corresponding picture using the *Boolean option*.

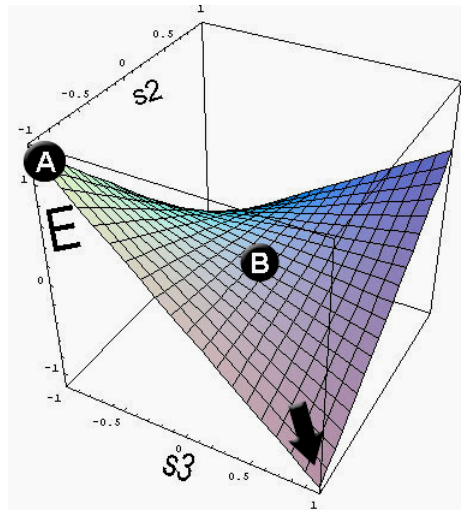


Figure 5: Energy landscape for calculating the asymptotic updates of $\langle 1 \ 0 \ 0 \rangle$. Starting points for energy minimization: **A** for the Boolean option, **B** for the DeMorgan option.

By beginning near the centre of the cube (B) and searching using gradient descent, the network has better chances of finding a global minimum than by beginning on the top position A. Hence, the De Morgan option bears a real advantage of improving the performance of the system. In Hopfield and Tank

networks (Hopfield & Tank, 1985) this advantage is regularly exploited, and the preferred option is to start the search from the centre of the cube.

Another advantage of the De Morgan option concerns the conceptual simplicity and naturalness of solving the extraction problem. Of especial importance is that the thresholds can be assumed to be zero in cases where the De Morgan option is chosen (with a working space $[-1, 1]$ of a unit). Hence, the additional term $-\sum_i \theta_i \cdot s_i$ can be dropped, which leads to a considerable simplification of the translation that transforms symmetric networks into symbolic knowledge bases.

A third advantage has to do with the explanation of recoverability (*bidirectionality*). In natural language theories this trait refers to a general characteristic of the form-meaning relation realized in understanding/production: *what we produce we are able to understand adequately and what we understand we are able to produce adequately*. Using the DeMorgan option of interpreting activation states, this picture will make the explanation much more transparent than the Boolean option.

In the abstract framework of pattern association patterns at a level A are associated with patterns at level B. Recoverability/bidirectionality can now be formulated as follows. We assume a simple experimental situation where a subject is presented with a (repeated) series of pairs $[a_i, b_i]$ of pattern from $A \times B$. The subject has to learn to produce the associated element, say b_i when the first member a_i of the pair is presented. Hence, in this paradigm the subject has to learn a predefined relation between a set of input patterns a_i and a set of output patterns b_i . For instance, an input pattern can be a lexigram (e.g. senseless syllable), and an output pattern can be a picture of a fruit. We assume a 1–1 correspondence between inputs and outputs.

If subjects are qualified to match stimulus a_i to b_i and then, without further training, match b_i to a_i , they have passed a **test of symmetry**. Passing this test, thus conforms to the characteristic of recoverability or bidirectionality in the domain of natural language computation. The test of symmetry plays an important role in research on the acquisition of functional symbol usage in apes and children. The important empirical finding is that children as young as 2 years pass the symmetry test (e.g. Green, 1990). In contrast, chimps did not show symmetry: having learned to match lexigram comparisons to object samples, the chimps were not able, without further training, to match the same objects now presented as comparisons to the corresponding lexigrams, now presented as samples (cf. Savage-Rumbaugh, 1984; Dugdale & Lowe, 2000).¹⁸

¹⁸ A possible exception is Kanzi, the bonobo monkey. Kanzi's knowledge was reciprocal. There was no need to teach her separately to produce and to comprehend (Savage-Rumbaugh & Lewin, 1994).

Using symmetric networks it is very simple to account for recoverability (passing the symmetry test) after learning the association $a_i \rightarrow b_i$ (assuming a 1-1 correspondence). For simplicity, we adopt a localist model with two levels of nodes such that the nodes correspond to the pattern a_i and b_i , respectively. Using the DeMorgan option, this corresponds to a system of weighted constraints $\{[a_i \leftrightarrow b_j: w_{ij}], 1 \leq i, j \leq N\}$ plus strict inhibitory links within the level A and B, respectively: $\{[a_i \leftrightarrow \neg a_j: \infty], i \neq j\} \cup \{[b_i \leftrightarrow \neg b_j: \infty], i \neq j\}$. Now it is not difficult to show that we can reproduce the list $a_i \rightarrow b_i$ for all i if and only if $w_{ii} > \sum_{1 \leq j \leq N, j \neq i} w_{ij}$ for each $1 \leq i \leq N$. That conforms to getting the inferences $a_i \approx_{PRK} b_i$ with the corresponding knowledge base PRK. Because of the symmetry of the knowledge base it can be concluded that the list can be reproduced in reverse order: $b_i \rightarrow a_i$ (i.e. $b_i \approx_{PRK} a_i$).

Concluding this section we can say that the DeMorgan option has a series of advantages if compared to the Boolean option: (i) it accounts to the idea of underspecification and inferential completion; (ii) it helps to improve the performance of the optimization procedure; (iii) it provides a conceptually simple and natural solution to the extraction problem; (iv) it makes the feature of recoverability transparent.

6 Optimality Theory and Symmetric Networks

Optimality theory (OT) was initiated by Prince & Smolensky (1993/2004) as a new phonological framework that deals with the interaction of violable constraints. In recent years, OT was the subject of lively interest also outside phonology. Students of morphology, syntax and natural language interpretation became sensitive to the opportunities and challenges of the new framework (e.g. Blutner & Zeevat, 2004). The reasons for linking scientists into this new research paradigm is manifold: (i) the aim to decrease the gap between competence and performance, (ii) interest in an architecture that is closer to neural networks than to the standard symbolist architecture, (iii) the aim to overcome the gap between probabilistic models of language and speech and the standard symbolic models, (iv) the logical problem of language acquisition, (v) the aim to integrate the synchronic with the diachronic view of language.

In the present context we emphasize the second motive. OT is deeply rooted in the connectionism paradigm of information processing. As a consequence, OT does not assume a strict distinction between representation and processing. The development of OT demonstrates a new and exciting research strategy: augmenting and modifying symbolist architecture by integrating

insights from connectionism. The development of Penalty Logic is another illustration of this strategy.

It's not possible to give a systematic introduction into OT here.¹⁹ The primary aim of this section is to draw attention to the close similarities between OT and the logical approach proposed in Sections 4 and 5, but also to point out some significant differences. The main difference between OT and numerical theories like *Penalty Logic* and *Harmonic Grammar* (Smolensky, 1986, 1995) is the shift from numerical to non-numerical constraint satisfaction. Why Prince and Smolensky (1993) proposed this shift, is explained by Smolensky (Smolensky, 1995: 266) as follows: “Phonological applications of Harmonic Grammar led Alan Prince and myself to a remarkable discovery: in a broad set of cases, at least, the relative strengths of constraints *need not be specified numerically*. For if the numerically weighted constraints needed in these cases are ranked from strongest to weakest, it turns out that each constraint is stronger than all the weaker constraints *combined*.” In other words, the shift from Harmonic Grammar to Optimality Theory, that means the realization of what is called *strict dominance of the OT constraints* appears to be mainly motivated by empirical findings in the domain of phonology.

A possible advantage of strict dominance lies in the robustness of processing. Following a suggestion of David Rumelhart the following argument was put forward: “Suppose it is important for communication that language processing computes global harmony maxima fairly reliably, so different speakers are not constantly computing idiosyncratic parses which are various local Harmony maxima. Then this puts a (meta-)constraint on the Harmony function: it must be such that local maximization algorithms give global maxima with reasonably high probability. Strict domination of grammatical constraints appears to satisfy this (meta-)constraint.” (Smolensky 1995, note 38: 286).

In concord with this argument it is not implausible to assume that the theoretical explanation for differences between automatic and controlled psychological processes (Schneider & Shiffrin, 1977) can also be seen as an emergent effect of the underlying neural computations (cf. Blutner, 2004). Whereas controlled processing relates to the capacity-limited processing when the global harmony maxima (= global energy minima) are difficult to grasp, automatic processing relates to a mode of processing where most local harmony maxima are global ones.

In order to illustrate the strictness of domination of grammatical constraints I consider a small fragment of the vowel system of English (cf. Kean, 1995),

¹⁹ For good introductions the reader is referred to the literature (e.g. Archangeli & Langendoen, 1997; Kager, 1999; Smolensky & Legendre, to appear).

which is roughly simplified for the present purpose.²⁰ The example rests on a classification of the vowels in terms of the binary phonemic features as illustrated in Table 4.

Table 4: Fragment of the vowel system of English and the phonological feature specifications

	/a/	/i/	/o/	/u/	/ɔ/	/e/	/æ/
<i>back</i>	+	—	+	+	+	—	—
<i>low</i>	+	—	—	—	+	—	+
<i>high</i>	—	+	—	+	—	—	—
<i>round</i>	—	—	+	+	+	—	—

For the purpose of applying propositional Penalty Logic, the phonological features may be represented by the atomic symbols BACK, LOW, HIGH, ROUND. The knowledge of the phonological agent concerning this fragment may be represented by the following violable constraints (usually called *markedness conventions*)²¹:

- (45) a. $\text{VOC} \leftrightarrow \text{BACK} \quad \varepsilon^1$
b. $\text{BACK} \leftrightarrow \text{LOW} \quad \varepsilon^2$
c. $\text{BACK} \leftrightarrow \sim \text{HIGH} \quad \varepsilon^3$
d. $\text{LOW} \leftrightarrow \sim \text{ROUND} \quad \varepsilon^4$

With regard to the agent's knowledge, the feature specifications in Table 4 are highly redundant. It can be shown that only the feature specifications in the grey fields must be given, the specification in the remaining fields can be calculated by the agent's knowledge. For the proper working of the constraint system in (45) it is required that the constraints are ordered in a hierarchical way, with (45a) at top and (45d) at bottom. This hierarchy corresponds to a relation of strict domination: one violation of a higher ordered constraint cannot be overpowered by arbitrary many violations of lower ordered constraints. The technical means of expressing the hierarchy is the use of *exponential penalties* with a basis $0 < \varepsilon \leq 0.5$. In the present case, $\varepsilon = \frac{1}{2}$ or smaller is a proper base since we are concerned with binary features which can be applied only once in each case.


²⁰ I borrow this example from Blutner (2004).

²¹ Further, two hard constraints are needed to express strong redundancies: $\text{LOW} \rightarrow \sim \text{HIGH}$; $\text{ROUND} \rightarrow \text{BACK}$.

Table 5 illustrates a sample calculation using an OT tableau. As usual in the OT literature a violation of a constraint is indicated by * and the *small hand* icon is used to mark the optimal candidate. In the present case we have only two candidates that satisfy the input's conditions for a non-high front vowel. The only free feature corresponds to \pm LOW. It resolves to –LOW because of the second constraint, which is the highest ranked constraint that discriminates the two candidates: it is satisfied for the optimal candidate but violated for the other candidate. The optimal candidate distinctively characterizes the vowel /e/. In the last column penalties are calculated from the constraint violations assuming penalties ε^n for constraints of rank n (with $\varepsilon = \frac{1}{2}$).

Table 5: OT tableau for calculating the optimal non-high front vowel: /e/

Input: +VOC \wedge –BACK \wedge –HIGH

	–	+	–	–	*	*	*	*	0.5550
	–	–	–	–	*		*	*	0.5055
BACK LOW HIGH ROUND	↓	↓	↓	↓	VOC	BACK	BACK	LOW	<i>Penalty</i> ($\varepsilon = \frac{1}{2}$)
					BACK	LOW	~HIGH	~ROUND	

Using the DeMorgan option it is straightforward to translate the constraint system (45) into a localist symmetric network as can be seen from Figure 6.

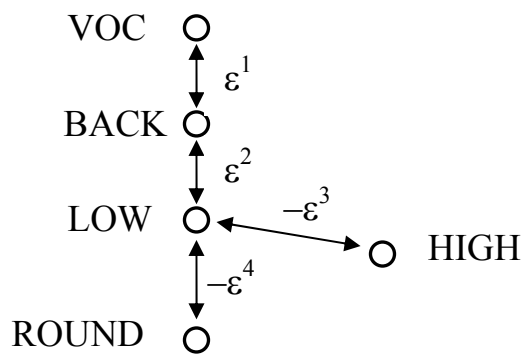


Figure 6: Hopfield network with exponential weights representing the generic knowledge²² of a phonological agent

²² The hard constraints mentioned in footnote 20 are not represented in this network. We leave it as an exercise for the interested reader to perform the corresponding modifications using the techniques explained in section 5.

We conclude that both Penalty Logic and Optimality Theory look for an optimal satisfaction of a system of conflicting constraints. Most importantly, the exponential form of the penalty function results in *strict domination* of the constraints, meaning that violations of many lower ranked constraints invariably count less than one violation of a higher ranked constraint. Moreover, we have seen how constraints that conform to formulae of propositional logic can be translated into a symmetric connectionist network by assuming a localist interpretation of the atomic symbols.

Early proposals to ground OT in connectionist architecture made use of (non-symmetric) feedforward networks (cf. Smolensky, 1986; Prince & Smolensky, 1993/2004). However, Smolensky & Legendre (to appear) also acknowledged the relevance and power of symmetric networks for developing an integrated connectionist/symbolic cognitive architecture. One important advantage of symmetric networks is that they give a natural account of the emergence of recoverability and bidirectionality.

There are two shortcomings with the presented account of reducing OT to connectionist networks. The first one concerns the use of a localist interpretation. Though a localist interpretation generates a fairly transparent relationship between symbols and node activities, this idea is much too naïve to be taken seriously as a promising programme in Cognitive Neuroscience. In realistic examples, the relation between the symbolic expressions as used in Penalty Logic and the elements of the pre-symbolic product space is much less direct than localist Hopfield models suggest. In an outstanding dissertation, Martinez (2004) proposed ideas for simultaneously using discrete symbolic means and non-discrete numerical means, and she developed tools of relating the two different realms in a much less direct way than strictly localist accounts suggest (see also Barwise & Seligman, 1997; Martinez, 2003). I think these ideas have a big potential for future accounts for an integrated connectionist/symbolic cognitive architecture.

The second shortcoming relates to the fact that the constraints we used in the example from intrasegmental phonology are *micro-constraints* in the sense that they are in direct correspondence to a very small fragment of the network. In fact, in the case under discussion each constraint corresponds to a pair of two linked nodes in the network. It is also indispensable to have constraints that correspond to larger parts of a network even when a localist interpretation is used. The whole idea of assemblies we mentioned in section 2 suggests that constraints are distributed over significant parts of the network. Hence, it is opportune to propose an extended scheme. In this connection I will introduce the notion of *macro-constraints*. In a first approximation, macro-constraints can be

seen as an organized congregation of micro-constraints, and they can be considered to constitute *innate structure*. The idea of macro-constraints is closely related to the idea of an *abstract genome* as developed by Smolensky & Legendre (to appear: Chapter 21). In detail, the idea has been worked out for basic CV syllable theory.

Macro-constraints can be defined as collections of micro-constraints with identical penalties. The idea of associating micro-constraints with identical penalties becomes appealing when we translate the set of micro-constraints into a neural net. Then identical penalties correspond to fixed relationships between certain connection weights in the symmetric network. For instance, let's assume a weighted macro-constraint $\mathbf{C} = \{p_i \leftrightarrow p_j, i \neq j\} : w$, where w is the penalty associated with all the micro-constraints $p_i \leftrightarrow p_j$ in \mathbf{C} . Hence, all weights between the nodes i and j in the network are required to be identical and to have the value w . Though the penalties can be changed by learning it is assumed that the identity of the corresponding weights is not lost over the course of learning. Thus this relationship is maintained during learning, although the absolute magnitude of the weights changes as particular knowledge is acquired. As a consequence, the relationship between connection weights can be considered to constitute the innate knowledge provided by a constraint (cf. Smolensky & Legendre, to appear).

Concluding, macro-constraints are essential for two related reasons: (i) they correspond to larger parts of the network and constitute assemblies, (ii) they express an innate relationship, which is not influenced by learning.

In sections 4 and 5 we have formalized a penalty (penalty/reward) knowledge base as a triple $\langle At, \Delta, k \rangle$ where Δ was a system of propositional expressions (=micro-constraints). Now we consider macro-constraints as (non-empty) sets of micro-constraints, and a macro-knowledge base MK can be defined as a corresponding triple $\langle At, {}^M\Delta, {}^Mk \rangle$, where (i) ${}^M\Delta$ is a set of nonempty sets of consistent sentences built on the basis of At ; (ii) ${}^Mk: {}^M\Delta \Rightarrow (0, \infty)$, the penalty function that associates penalties with each macro-constraint. Now the system energy of an interpretation v with regard to a macro-knowledge base MK is defined as follows:

$$(46) \quad \mathcal{E}_{MK}(v) =_{\text{def}} -\sum_{\mu \in {}^M\Delta} {}^Mk(\mu) \sum_{\delta \in \mu} \llbracket \delta \rrbracket_v$$

For each macro-knowledge base $MK = \langle At, {}^M\Delta, {}^Mk \rangle$ we can construct the associated ordinary knowledge base $K = \langle At, \Delta, k \rangle$, where $\Delta = \cup {}^M\Delta$ and $k(\delta) = {}^Mk(\mu)$ if $\delta \in \mu$. It is obvious that the system energy (47) of an interpretation with regard to a macro-knowledge base is identical to the system energy of an interpretation with regard to the associated ordinary knowledge base: $\mathcal{E}_{MK}(v) = \mathcal{E}_K(v)$. The crucial point is that the penalties $k(\delta)$ for all micro-constraints δ that

constitute the macro-constraint μ are identical. Further, it is obvious how to construct the symmetric network that corresponds to a macro-knowledge base: build the associated ordinary (micro-) knowledge base and translate it into the network using the technique explained in sections 4 and 5.

7 Conclusions: Logic and embodied theories of cognition

The present contribution can be seen as part of recent efforts to develop an embodied view of cognition. The emerging viewpoint of embodied cognition holds that cognitive processes are deeply rooted in the body's interactions with the world (cf. Brooks (1999); Anderson (2003); Lakoff & Johnson (1999); Varela, Thompson, & Rosch (1993)). The idea of embodiment has diverse aspects. Several philosophers and cognitive scientist agree that at least the following three aspects are of special importance (cf. Anderson, 2003):

- Reductionist aspect: The system must be realised in a coherent, integral physical/biological structure. As an immediate consequence, certain features of the symbolic system (e.g. the OT Grammar) must be reducible to plausible neural models.
- Evolutionary aspect: The explanation of the behaviour must include reference to cultural evolution. This derives from the observation that intelligence lies less in the individual brain and more in the dynamic interaction of brains with the wider world, including especially the social and cultural worlds.²³
- Grounding aspect: Symbol-manipulation has to be grounded in non-symbolic function. OT constraints are embodied, not disembodied. A symbol is grounded if it has its meaning or content by virtue of its causal properties and relations to the referent of the symbol. Hence, symbols have to be grounded ultimately in the sensory-motor system or other bodily systems or are appropriately defined in terms of grounded symbols.

The research program of embodied cognition is a continuation of the program of *situated* cognition. It is the centrality of the *physical grounding project* in

²³ In the domain of linguistics, Jackendoff (2002) makes the following remarkable claim stressing the influence of cultural interaction in understanding language: "If some aspects of linguistic behaviour can be predicted from more general considerations of the dynamics of communication in a community, rather than from the linguistic capacities of individual speakers, then they should be." (Jackendoff 2002:101).

embodied cognition that differentiates these two research programs (cf. Anderson, 2003).

Taking up the view of embodiment, the present article builds mainly around the reductionist aspect of embodiment. What are the central general principles of computation in connectionist – abstract neural – networks? How can these principles be reconciled with those of symbolic computation? Which basic assumptions of OT can be reduced to connectionist computation? And in what case alternate explanations are required? In a nutshell, we can state the following main results:

- To overcome the gap between symbolism and connectionism it is useful to view symbolism as a high-level description of the properties of (a class of) neural networks. The application of algebraic and model-theoretic techniques for a higher-level analysis of neural networks (e.g. Balkenius & Gärdenfors, 1991; Pinkas, 1995; Blutner, 1997, 2004) and their development in the present paper proves especially valuable when it comes to study the concrete link between inferences in symmetric networks and inferences in nonmonotonic logic.
- The foundational issue of OT: The general shape of symbolic OT systems proves to be conforming to the penalty-logical treatment proposed in sections 4 and 5. Because of the close relations between Penalty Logic and symmetric networks, certain features of standard OT appear to be reducible to the basic traits of neural network models. This concern first at all the idea of domination: constraint conflict is resolved via a notion of differential strength: stronger constraints prevail over weaker ones in cases of conflict.
- Strictness of domination (hierarchical encoding of constraint strengths): This problem matters both from a theoretical and an empirical perspective. In the words of Bechtel, the solution to this problem “may create a rapprochement between network models and symbolic accounts that triggers an era of dramatic progress in which alignments are found and used all the way from the neural level to the cognitive/linguistic level (Bechtel, 2002, p.17). Presently, there are only vague ideas about how to account for the strictness of domination and the entailed idea that Grammar (usually) does not count. Moreover, it is rather unclear how to give a theoretically satisfying account for explaining under which conditions the strict domination of constraints applies and under which conditions it does not.
- The idea of macro-constraints is essential for matching larger parts of a network (assembly formation). Further, macro-constraints can be used to express innate relationships on symmetric networks – i.e. relationships that aren't controlled by learning.

Standard OT respects the generative legacy in assuming that the universal features of language can be explained by assuming a Universal Grammar (UG). UG describes the innate knowledge of language that is shared by individual humans. In standard OT, the innate knowledge of language consists (a) of a generative device that generates the admissible input-output pairs and (b) the set of constraints. Language-particular aspects refer to the possible rankings of the constraints (e.g. Prince & Smolensky, 1993/2004). Hence, the suggestion of an abstract genome (Smolensky & Legendre, to appear) as well as the suggestion of macro-constraints and the way they constrain symmetric networks nicely fits into this picture.

However, recent effort on the problem of the evolution of language in humans (e.g. Hurford, 1998; Steels, 1998; Kirby, 2002; Zeevat & Jäger, 2002) made clear that a thorough explanation of the universal properties of language cannot be exclusively based on an individual's cognitive capacity which is taken to be biologically determined. So, if we want to know how and where the universal features of language are specified, it is not sufficient to consider only an individual's competence and how it is derived from primary linguistic data via the Language Acquisition Device (LAD). Rather, it is essential to focus on how certain hallmarks of human language can arise in the absence of biological change by assuming the force of *cultural evolution*. In explaining the universal properties of language, the evolutionary approach is in line with the claims made by proponents of embodied cognitive science. Hence, it is our central task to investigate the interaction between biological and cultural substrates. The paradigm of iterated learning (e.g. Kirby & Hurford, 1997; Kirby, 2002) has proven as especially useful in investigating the emerging effects from this interaction.

Taking the dimension of cultural evolutionary into account suggest that at least some principles of OT can be explained as emergent factors of cultural exchange. This concerns, first at all the explanation of bias constraints (Zeevat & Jäger, 2002) and the principle of constructional iconicity²⁴, which is related to the feature of weak bidirection (Mattausch, 2004). Hence, naïve OT with its assumption of inborn constraints has to be overcome by an embodied OT, which respects the role of grounding constraints by iterated learning. In this regard it is important that the mechanism of grounding is directed by mechanisms that are

²⁴ Constructional iconicity states that there is a harmonic linking between complex semantic contents and complex (surface) forms on the one hand and less complex semantic contents and simple forms on the other hand. Both in pragmatics and in (natural) morphology the principle plays an important role in describing the *direction of language change*. In formal semantics, this principle is called *division of pragmatic labour* (Horn, 1984); in the school of „natural morphology“ it is called *constructional iconicity* (Wurzel, 1998).

very close to those used in modelling evolutionary change (e.g. Hayes, 1996; Boersma, 1998).

In this article I have concentrated on the reductionist aspect of embodied cognition – certain features of a symbolic system (e.g. the OT Grammar) must be reducible to plausible neural models. Though the reductionist programme is an integral part of the embodied paradigm it is not the whole story. The evolutionary aspect and the aspect of grounding likewise deserve attention. Once more, the feature of situatedness, i.e. dynamic interaction of brains with the wider world, including especially the social and cultural worlds, should prove promising for future research.

8 References

- Anderson, M. L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence*, 149, 91–130.
- Archangeli, D., & Langendoen, D. T. (1997). *Optimality theory: An overview*. Malden, MA/Oxford, UK: Blackwell.
- Asimov, I. (1950). *I, Robot*: Gnome Press.
- Balkenius, C., & Gärdenfors, P. (1991). Nonmonotonic inferences in neural networks. In J. A. Allen & R. Fikes & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning*. San Mateo, CA: Morgan Kaufmann.
- Barwise, J., & Seligman, J. (1997). *Information flow: the logic of distributed systems*. New York: Cambridge University Press.
- Bechtel, W. (2002). *Connectionism and the Mind*. Oxford: Blackwell.
- Blutner, R. (1997). Nonmonotonic logic and neural networks. In P. Dekker & M. Stokhof & Y. Venema (Eds.), *Proceedings of the eleventh Amsterdam Colloquium* (pp. 79-84): ILLC/Department of Philosophy, University of Amsterdam.
- Blutner, R. (2004). Nonmonotonic inferences and neural networks. *Synthese (Special issue Knowledge, Rationality and Action)*, 142, 143-174.
- Blutner, R., & Zeevat, H. (Eds.). (2004). *Optimality Theory and Pragmatics*. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.
- Boersma, P. (1998). *Functional phonology*. The Hague: Holland Academic Graphics.
- Brooks, R. (1999). *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.
- Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, MA: MIT Press.
- Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, 815-826.

- d'Avila Garcez, A. S., Broda, K., & Gabbay, D. M. (2001). Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence*, 125, 155-207.
- deSaint-Cyr, F. D., Lang, J., & Schiex, T. (1994). Penalty logic and its link with Dempster-Shafer theory, *Proceedings of the 10th Int. Conf. on Uncertainty in Artificial Intelligence (UAI'94)* (pp. 204-211).
- Dugdale, N., & Lowe, C. F. (2000). Testing for symmetry in the conditional discriminations of language-trained chimpanzees. *Journal of the Experimental Analysis of Behavior*, 73, 5-22.
- Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, 6, 205-254.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28, 3-71.
- Gerstner, W., & Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, Mass.: Cambridge University Press.
- Graben, P. b. (2004). Incompatible Implementations of Physical Symbol Systems. *Mind and Matter*, 2, 29-51.
- Green, G. (1990). Differences in development of visual and auditory-visual equivalence relations. *Journal of the Experimental Analysis of Behavior*, 51, 385-392.
- Hajek, P. (1998). *Metamathematics of fuzzy logic*. Dordrecht: Kluwer.
- Hayes, B. P. (1996). Phonetically Driven Phonology: The Role of Optimality Theory and Inductive Grounding.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.
- Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference, *Proceedings of the Institute of Electronic and Electrical Engineers Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 448-453). Washington, DC: IEEE.
- Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzman machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I and II*. (pp. 282-317). Cambridge, MA: The MIT Press/Bradford Books.
- Hölldobler, S. (1991). Towards a connectionist inference system. In N. Cercone & F. Gardin & G. Valle (Eds.), *Computational Intelligence III* (pp. 25-38): North-Holland.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79, 2554-2558.
- Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci.*, 81, 3088-3092.
- Hopfield, J. J., & Tank, D. W. (1985). Neural computation of decisions in optimization problems. *Biol. Cybern.*, 52, 144-152.

- Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.
- Hurford, J. R. (1998). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77, 187–222.
- Jackendoff, R. (2002). *Foundations of Language*. Oxford: Oxford University Press.
- Jibu, M., & Yasue, K. (1995). *Quantum Brain Dynamics and Consciousness*. Amsterdam/Philadelphia: John Benjamins.
- Kager, R. (1999). *Optimality theory*. Cambridge: Cambridge University Press.
- Kean, M. L. (1995). *The theory of markedness in generative grammar*. Unpublished Ph.D. thesis, MIT, Cambridge, Mass.
- Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8, 185–215.
- Kirby, S., & Hurford, J. (1997). *The evolution of incremental learning: language, development and critical periods*. Edinburgh: University of Edinburgh.
- Kokinov, B. (1997). Micro-level hybridization in the cognitive architecture DUAL. In R. Sun & F. Alexander (Eds.), *Connectionist-symbolic integration: From unified to hybrid approaches* (pp. 197-208). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.
- Maass, W. (1999). *Pulsed Neural Networks*. Cambridge, Mass.: MIT Press.
- Martinez, M. (2003). Towards a model of heterogeneous commonsense reasoning. In J. Baldwin & R. d. Queiroz & E. H. Hauesler (Eds.), *Proceedings of WoLLIC'2001. Matematica Contemporanea, V 24*: Sociedade Brasileira de Matematica.
- Martinez, M. (2004). *Commonsense reasoning via product state spaces*. Unpublished PhD, Indiana University, Bloomington.
- Mattausch, J. (2004). *On the Optimization & Grammaticalization of Anaphora*. Unpublished Ph.D. Thesis, Humboldt University, Berlin.
- McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I and II*. Cambridge, MA: The MIT Press/Bradford Books.
- Palm, G., & Wennekers, T. (1997). Synchronicity and its use in the brain. *Behavioral and Brain Sciences*, 20, 295-296.
- Pinkas, G. (1992). *Logical inference in symmetric connectionist networks*. Unpublished Doctoral thesis, Washington University, St Louis, Missouri.
- Pinkas, G. (1995). Reasoning, connectionist nonmonotonicity and learning in networks that capture propositional knowledge. *Artificial Intelligence*, 77, 203-247.
- Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, 36, 27-47.
- Pribram, K. H. (1991). *Brain and Perception*. New Jersey: Lawrence Erlbaum.

- Prince, A., & Smolensky, P. (1993). *Optimality theory*. Rutgers Center for Cognitive Science: Technical Report RuCCSTR-2.
- Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Rutgers University and University of Colorado at Boulder: Technical Report RuCCSTR-2, available as ROA 537-0802. Revised version published by Blackwell, 2004.
- Rojas, R. (1996). *Neural Networks - A Systematic Introduction*. Berlin, New-York: Springer-Verlag.
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In J. L. McClelland & D. E. Rumelhart & the-PDP-Research-Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition I*. Cambridge, MA: MIT Press.
- Savage-Rumbaugh, E. S. (1984). Acquisition of functional symbol usage in apes and children. In H. L. Roitblat & T. G. Bever & H. S. Terrace (Eds.), *Animal Cognition* (pp. 291-310). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Savage-Rumbaugh, S., & Lewin, R. (1994). *Kanzi : The Ape at the Brink of the Human Mind*. John Wiley & Sons.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing. *Psychological Review*, 84, 1-66.
- Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3, 417-457.
- Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rule, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16, 417-494.
- Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypotheses. *Ann. Rev. Neuroscience*, 18, 555-586.
- Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. volume 1: Foundations* (pp. 194-281). Cambridge, Mass., London: MIT.
- Smolensky, P. (1995). Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald & G. Macdonald (Eds.), *Connectionism: Debates on Psychological Explanation* (pp. 221-290). Oxford: Blackwell.
- Smolensky, P., & Legendre, G. (to appear). *The Harmonic Mind: From neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.
- Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103, 133-156.
- Varela, F. J., Thompson, E., & Rosch, E. (1993). *The Embodied Mind*.

- von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Report 81-2). Göttingen: Max-Planck-Institut für Biophysikalische Chemie.
- Wennekers, T. (1999). *Synchronisation und Assoziation in neuronalen Netzen*. Aachen: Shaker Verlag.
- Wennekers, T., & Palm, G. (2000). Cell assemblies, associative memory and temporal structure in brain signals. In R. Miller (Ed.), *Time and the Brain. Conceptual Advances in Brain Research, vol II*: Harwood Academic Publishers.
- Wurzel, W. U. (1998). On markedness. *Theoretical Linguistics*, 24, 53-71.
- Zeevat, H., & Jäger, G. (2002). *A statistical reinterpretation of harmonic alignment*. Paper presented at the 4th Tbilisi Symposium on Logic, Language and Linguistics, Tbilisi.

Signalling Games: Evolutionary Convergence on Optimality¹

Tom Lentz

Utrecht Institute of Linguistics OTS, Utrecht University

Reinhard Blutner

ILLC, University of Amsterdam

Horn's *division of pragmatic labour* (Horn, 1984) is a universal property of language, and amounts to the pairing of simple meanings to simple forms, and deviant meanings to complex forms. This division makes sense, but a community of language users that do not know it makes sense will still develop it after a while, because it gives optimal communication at minimal costs. This property of the division of pragmatic labour is shown by formalising it and applying it to a simple form of *signalling games*, which allows computer simulations to corroborate intuitions. The division of pragmatic labour is a stable communicative strategy that a population of communicating agents will converge on, and it cannot be replaced by alternative strategies once it is in place.

1 Introduction: philosophy and empiricism

If philosophy is the justification of knowledge, one of the subjects that may be justified is empiricism as a source of knowledge. But, reversely, can empiricism justify philosophical principles? Our research is based on simulation, a form of empiricism, to test the hitherto unproven but plausible evolutionary origin of a theory of language philosophy, namely Horn's *division of pragmatic labour*.

An important aspect of linguistic theory should be the possibility to account for the origin and development of language. Some language universals might be

¹ We thank Erik Borra, Arnold Obdeijn, Jasper Uijlings, and Reinier Zevenhuijzen who contributed to a very early version of this paper written in Dutch, and to the actual computer simulations on which this article is based. The first author is supported by a grant from the Netherlands Organisation for Scientific Research (NWO), project grant 277-70-001 to René Kager.

explained by, or reduced to, properties of the brain, whether these are general cognitive properties or specific to language. As the brain can be observed, theories positing properties of the brain are, at least in theory, empirically testable.

Historical experimentation or observation, on the other hand, is virtually impossible, making it very hard to test theories on language universals and their origins. This paper, however, uses simulations of language development to overcome this problem. The universal property that the paper focuses on is a pragmatic property of language, Horn's *division of pragmatic labour* (Horn, 1984) which will be explained below. This phenomenon is described as a property that is observed universally in language use; however, it is posited philosophically as a basic principle that is used to describe natural language semantics and pragmatics. This paper attempts to add more arguments to the philosophical side of the principle by explaining its emergence as virtually inevitable under reasonable assumptions on language evolution.

It is assumed that language came into being by linking signals to meanings and vice versa². Given this assumption, this article is to show that the division of pragmatic labour follows from repeated acts of linguistic communication. This is important, as a population without language cannot agree on how to develop a language. This makes it undesirable to attribute pragmatic preferences of language use to individual preferences in language users, as individuals have no reliable information on the preferences of other language users a priori. Computer simulations showed that it is not necessary to assume individual preferences to be biased towards optimal pragmatic solutions; the only 'bias' should be that effective communication is preferred over ineffective communication, but this 'bias' cannot be seen as a property that determines the strategies of individual language users, as it surpasses the level of the individual. Still, an optimal solution emerges that happens to conform to the division of pragmatic labour.

In this paper, an initial stage of unprincipled form-meaning pairings is assumed. An individual may produce a specific signal (noise) in a specific situation. This signal is his expression of that situation. However, this "language" is restricted to the level of the individual and it is quite removed from a shared communicative device.

In the next section, we explain what is meant by a Horn strategy of form-meaning pairings. Section 3 introduces Lewis' idea of a signalling game (Lewis, 1969). Subsequently, in section 4 we explain our implementation of a simulation experiment of signalling games, and in section 5 and 6 we present our results

² In the modern constructionist literature, such form-meaning pairs are called constructions (see Goldberg, 1995; Tomasello, 2003).

demonstrating how Horn's division of pragmatic labour emerges from an evolutionary mechanism in language use. Further, we discuss the evolutionary stability of certain form-meaning pairs. Section 7, finally provides a general discussion including an outlook of how the present research could be continued.

2 Horn strategy

People tend to use the simplest signals for the most common messages and more complex signals for more unusual messages. This can be seen in example (1) , below.

- (1) a. John sings a song.
- b. John produces noises resembling a song.

Sentence (1a) caters for a normal situation where someone sings a song. In the second sentence something strange seems to happen. Indeed, why use a strange sentence for a message that may be catered for with a normal sentence?

This observation is expressed by the following rule: *Simple messages express normal situations and more complex messages express strange situations*. This rule was posited by Horn (1984) and is known as *Horn's division of pragmatic labour* or as *Horn's rule* or (here) as the *Horn strategy*.

Horn justifies his rule on empirical grounds, by observation. The lack of observations on its development, however, makes it hard to explain its origin. As theories gain empiric backing by a multitude of observations complying with it, we will continue this paper with a formalisation of Horn's rule that allows for observations on its development. Does Horn's rule always apply to the development of language or is it an accidental feature of the languages that we happen to observe? To answer this question, populations developing a language were simulated. In the simulations, language users (agents) interact to convey meaning in so-called signalling games. The language users are assumed to be defined by a genetic make-up, which can change over generations.

3 Signalling Games

The concept of *signalling games* (Lewis, 1969) can be summarised as follows: a population of agents communicate to each other; if they manage to interpret a message correctly, they score in the game. The signalling games paradigm is well-defined and therefore it can be put to use to simulate language development in a straight-forward way, with the addition of an evolutionary perspective.

Methodologically, the development of an evolutionary perspective can proceed in two different ways. First, there is the purely theoretical approach, by using the concept of an evolutionary stable strategy (Smith, 1982). For several interesting results that were found in this way, we refer the reader to van Rooy (2004) and Benz, Jäger, and van Rooy (2005). Second, there is the construction of explicit dynamic models of the process by which the proportions of various strategies in a population change. This approach was pioneered by Luc Steels (e.g. Steels, 1998; Steels & Belpaeme, 2004).

In this paper, we follow the second approach and we will develop a genetic algorithm that implements a signalling game, adding (to the general model) the idea that well communicating agents score points and are thus more likely to procreate.

In the model, there are two roles for every agent: sender and receiver. A sender sends a message covering the meaning he wants to transmit. The receiver interprets the message; he attributes a meaning to it. Prerequisite for good communication is that the meaning is the same in both cases; only then agents understand each other. Note that no a priori form-meaning relation is imposed. This is important, as even though one might accept the emergence of simple form-meaning pairs that have iconic value (like the imitation of an animal's call to signify that animal), assuming iconic "words" is in no trivial way sufficient to explain full-fledged languages. In addition, it also fails to explain Horn's division of pragmatic labour as the connection between simple form and common meaning is not iconic. To the human observer, that relation might be "logical", but that can be explained wholly by the fact that this is what we observe and/or have acquired, and therefore begs the question.

It is important to remark that communicating and procreating agents should not be seen as models of actual humans; the agents are way too simple and there is no evidence as of yet to indicate that pragmatic strategies are encoded directly in the genome; without any biological underpinning, such an assumption is therefore far too speculative. The simulations are meant to illustrate the high likelihood that pragmatic strategies, when adapted to communicative needs, converge on Horn-like states, thereby creating both a shared language without prior conference and without individual properties.

4 Formalisation of signalling games

4.1 Evolving agents in the signalling game

In the simulations, agents obtain points for each message they rightly transmit or interpret. Those scoring most points are most likely to survive and procreate. An agent is fully defined by his communication strategy. His only task is to

communicate and that is the sole thing that matters for his survival and procreation. The game is bilateral; each agent both sends and receives.

Offspring of two agents will have a communication strategy combining both strategies. This might be the origin of a common language, but that is merely accidental and only holds for children with similar parents, not in general. Obviously, this simplification of procreation does not model human children of which the parents speak a different language; a child raised in a bilingual situation is most likely to speak both languages. The offspring of agents with different strategies combine the strategies of their parents; one could also think of this as a probabilistic simplification, as the strategies are independent of each other.

4.2 Domain of communicative interaction

The game itself is a simplification of human communication. Agents may find themselves in two different situations, the *normal situation* and the *deviant situation*. In the model, this amounts to a desire of the agent to express the *normal meaning* and the *deviant meaning*. Agents may transmit two different signals, the *simple signal* and the *complex signal*. The names for the situations and for the meanings are mere labels; they have no properties that are different, the only difference is that they are not the same instance of the same class.

Each combination of meaning and signal is allowed. This leads to 2 (meanings) \times 2 (signals) = 4 possible sending strategies as well as 4 receiving strategies. Together this leads to 4 (sending strategies) \times 4 (receiving strategies) = 16 communication strategies.

The game is played as follows: two agents meet and communicate. One starts to speak. First his situation is normal, so he communicates the normal meaning in the form relevant to his strategy. The second agent interprets this signal with the meaning relevant to his strategy. If indeed this is the intended, normal meaning, both score one point. Next, the sender communicates the deviant meaning and the same happens. Subsequently, the roles are inverted. In this way both agents can score up to four points per game.

4.3 Representation of the agents

An agent is represented by a bitstring (a series of bits, i.e. zeros and ones). This bitstring represents his communication strategy. The bitstring of an agent with the Horn strategy, for instance, is 0101. The bits are 0 for simple and for normal and 1 for complex and for deviant and the positions indicate the following:

1. the signal used for the normal meaning (simple, 0, or complex, 1)

2. the signal used for the deviant meaning (simple, 0, or complex, 1)
3. the meaning attributed to the simple signal (normal, 0, or deviant, 1)
4. the meaning attributed to the complex signal (normal, 0, or deviant, 1)

These bitstrings may be considered to be the agent's genome. The bitstring can be transformed in a more insightful graph, as shown in Figure 1.

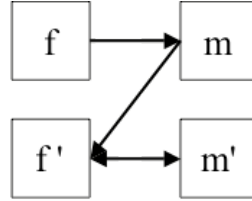


Figure 1: Example of an agent type. The arrows indicate the connections from forms to interpretations and from interpretations to forms, respectively

The total set of sixteen strategies is shown in Figure 2 in a schematic form.

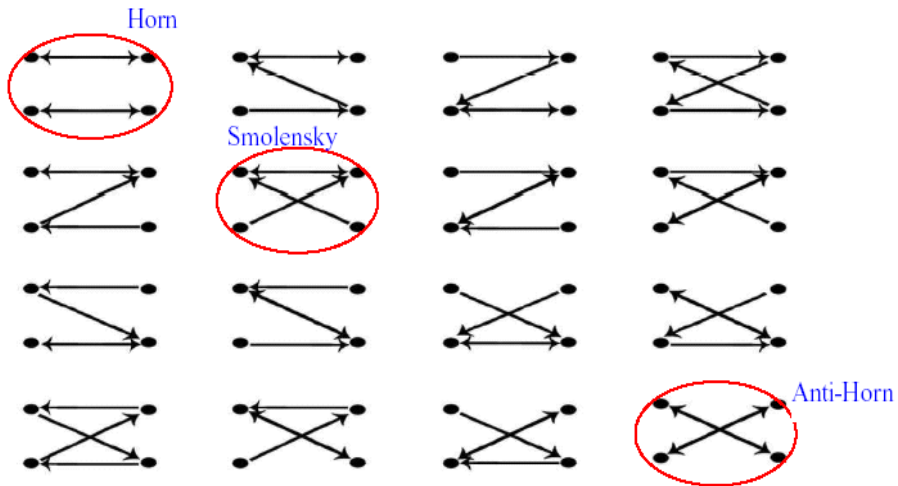


Figure 2: All communication strategies (agent types) in a schematic form. Three strategies are of special interest: (a) the Horn strategy as explained in section 3, (b) the Smolensky strategy reflecting the initial state of a learner (everything is assumed to be simple), (c) the anti-Horn strategy, which can be seen as the complement of the Horn strategy.

4.4 The selection mechanism

A selection procedure always chooses agents who survive and agents who procreate. The probability to be selected is based on the agent's number of points, it cannot be excluded that an agent with few points survives and procreates, though the chances are slim.

The number of points scored by two agents a and b is determined by adding up the two meanings i (i.e. the normal and deviant ones), as shown below:

$$\text{Points}(a) = \text{Points}(b) = \sum_i \delta[H_b(S_a(i)) , i] + \delta[H_a(S_b(i)) , i],$$

$$\text{where} \quad \delta(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

$H_x(f)$ gives the interpretation by agent x for a signal f .

$S_x(m)$ gives the signal used for it by agent x for a meaning m .

The following mechanism of *procreation* has been used: if two agents are selected as having offspring, a random point in the bitstring is chosen. Two new agents, children of two parents, are created. For the one child, the bits to the left of the point originate from the first parent, the bits to the right from the second parent. For the other child, the bits to the right of the point originate from the first parent, the bits to the left from the second parent.

At birth, agents might undergo a *mutation*. Every bit has a tiny chance to mutate after having been determined by the characteristics of its parents. This probability, the *mutation ratio*, was set to be 0.01 by default.

Convergence of a population is related to a strategy being predominant within a population, as it tends to be advantageous for agents to use that strategy, since they tend to be well understood and they do understand well. The same holds for humans: in England it is useful to speak English, because many people will understand it; in Japan it is better to speak Japanese. The utility of a language is not a function of the language alone, but of the population and the language in interaction.

Agents using the predominant strategy will have more offspring, making the population converge towards that strategy. However, not all strategies make communication optimal in a homogeneous population; in fact only two strategies allow to get the maximum of 400 points, namely the Horn and the anti-Horn strategy. Horn-agents are those agents that communicate the normal

meaning with the simple form, the deviant meaning with the complex form, and *vice versa* for interpretation (see Figure 2). Anti-Horn is the opposite.

As convergence does not necessary imply that a population is communicating in an optimal way, convergence and stability are defined as two separate concepts. Convergence is characterized as follows:

Convergence of a population takes place if a considerable majority of that population shares the same strategy in the limit.

The percentage constituting a considerable majority may be adjusted. Stability is informally characterized as follows:

A strategy is stable if an already existing majority of users of that strategy cannot be overwhelmed by another strategy, i.e. if the majority uses a strategy that is the best given the population.

This means that convergence towards a stable strategy is irreversible. In our evolutionary game a stable strategy may be compared to a Nash equilibrium. The Nash equilibrium is a concept from game theory developed by the mathematician John Nash (Nash, 1950). Two players find themselves in a Nash equilibrium if none of them gains by changing his behaviour. In our game it is not the interaction between two agents that matters, but the interaction between all agents. Moreover, the players themselves cannot change their behaviour. That means that in a way the entire population finds itself in equilibrium, a stable situation, with a certain strategy, if no strategy exists allowing an individual to score more points. This individual could come into existence by mutation or procreation; this is not likely to happen, because the mutation rate is low and in case a population is homogenous, children will usually be copies of their (identical) parents. However, if a different strategy is more successful, it can overtake the population given enough strategies, as it is more likely to procreate. In case of a Nash equilibrium, none of the fifteen possible alternative strategies is more successful in a population using the prevailing strategy.³

Finally, we should note that a stable strategy in the simulations amounts to a shared language; if a population converges to it, the language is shared.

³ For readers interested in precise definitions, we refer to Weibull (1995), van Rooy (2004), and Benz et al. (2005).

5 Simulations of signalling games strategy evolution

Given the above formalisations, a series of computer simulation was run to assess evolution and emergence of the signalling strategies. The simulations started with a start population of 100 agents. All agents interact (play the signalling game) with all other agents, and acquire points for successful communication, as described above. The selection mechanism is then applied to yield a new generation of agents; 85% of the agents perish, and are replaced by new agents, that are children of existing agents (cross-overs), with a 1% chance of mutation. All simulations were iterated 100 times, to assure findings were not due to chance.

5.1 Stability of Horn and anti-Horn

Procedure

In the first group of simulations, all agents had used the Horn strategy at the start of the evolution. Apart from that, the settings as described above are used.

Results

The Horn strategy is an excellent strategy throughout the evolution. The fitness of agents is close to 80 percent of the maximum (averaged over the 100 simulation experiments). The fitness percentage stays close to the maximum value, as most agents use the Horn strategy, even though mutation adds a low percentage of deviant strategies every generation. The population always converged towards Horn, in all 100 evolution simulations. The anti-Horn strategy never reaches dominance.

The mirror image emerges when the initial population is anti-Horn; in that case Horn never emerges as dominant strategy, and the population converges to anti-Horn in all of the 100 simulated evolutions.

Discussion

The fact that the Horn and anti-Horn strategies dominate populations that were initially already Horn respectively anti-Horn is not utterly surprising, but it is interesting to see that the result is so persistent. In all evolutions, Horn stays Horn and anti-Horn stays anti-Horn. This shows that the Nash equilibria that these two strategies exhibit, are very relevant to the evolution of the communication strategy. It can be explained why only these two strategies are stable. Communication between two Horn agents is the best possible and thus scores the maximum number of points. The same goes for two anti-Horn agents. All other strategies show weaknesses of fitness. One of the weaknesses is that

the send-strategy does not distinguish between normal and deviant meaning, if the send strategy is: “use the simple form for the normal meaning”, but at the same time: “use the simple form for the deviant meaning”. This strategy does not distinguish between situations. An agent using this strategy will never be understood in the best possible way, because there is no way of telling what meaning it tries to signal. Another weakness is when the send and receive strategies do not match. The send strategy is: “use the complex form for the normal meaning”. The receive strategy is “interpret the complex form as the deviant meaning”. Agents using such a strategy cannot communicate in the best possible way with identical others; if they want to communicate the normal meaning, the other agent will understand it as being deviant. However, this weakness only arises in. By chance, one agent with this strategy could perform very well, or even perfectly, if his strategies happen to go well with those of the other agents. However, the successful agent will have more offspring, filling the population with more and more agents with identical strategies; these agents cannot communicate very well. This effectively caps the total percentage of agents with inconsistent strategies; the cap is far below any reasonable convergence threshold, i.e., below 50%.

5.2 Starting simple or at random

Tesar and Smolensky (1998, 2000) describe a population of agents all using the simple strategy. In their terms, markedness is initial more important than faithfulness; difficult things are avoided. In the signalling game described above, it means that a simple signal is used for all meanings, and every signal is interpreted as a normal situation, a simple meaning. The idea is that a population will start with such a strategy, possibly because complex signals did not exist in former generations, and deviant meanings could not be distinguished from simple situations. If such a population would converge towards the Horn strategy, model and actual language would nicely match. Two series of 100 simulations were done to test the influence of the start population.

Procedure

To see whether the Nash-equilibriums strategies are indeed dominant, another series of hundred evolutions was simulated, now starting with the simplest strategy, the Smolensky strategy. That strategy is based on Smolensky’s idea that in acquiring language, simple forms and interpretations are preferred. Taking that to mean that no meaning is expressed with the complex form and no form is interpreted to be deviant, this conforms to the strategy highlighted in Figure 2 as Smolensky strategy. The whole procedure was repeated with mixed

strategies, in which agents in the start population have a random strategy, chosen from all sixteen possible strategies.

Results

In the evolutions starting with the Smolensky strategy, the Horn and anti-Horn strategies emerge, but neither of them seems to be able to tip the balance and dominate the population. As can be seen in Figure 3, either the Horn or the anti-Horn strategy manages to climb to slightly below 50%, but not further. This is usually due to the opposite strategy, or in a few cases a variant of that, that cannot improve by changing towards Horn or anti-Horn (all of these cases can one-by-one be explained, but the details are left out as they are not important for the general argument).

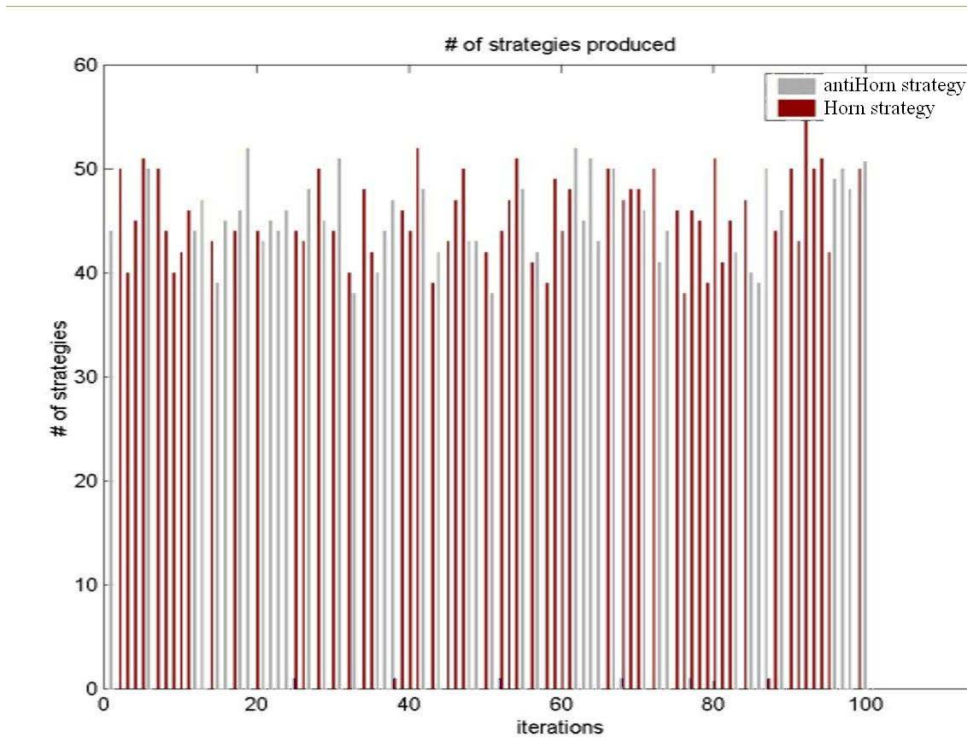


Figure 3: Percentage of agents using Horn or anti-Horn strategy at the end of 100 simulations (iterations) starting with homogenous Smolensky population.

Further, it was found that in simulations starting with mixed populations, convergence was almost perfectly divided amongst Horn and anti-Horn strategies.

Discussion

The fact that the Horn and anti-Horn strategies always emerge shows their evolutionary force. The fact that none of them is able to dominate a population shows that both are very stable, and could be compared to a population that is divided into two groups. No agent can communicate outside its group, but if its offspring changes towards the other group, it loses as much as it gains. Note that agents that do not communicate could still have offspring together, as procreation only depends on individual fitness scores. (It is up to the reader to assess if that is in accordance with real evolution.)

It is likely that mutations make some agents have offspring in the other group, but as this happens in both groups, neither of the groups is able to take advantage of that fact. This simulation, together with the above simulations, shows that evolutionary stability does not enforce an outcome, but that evolution can only go towards the two Nash-equilibriums.

6 The difference between Horn and anti-Horn

In the simulations described above, convergence towards both Horn and anti-Horn occurred (in a 50:50 ratio), whatever the start population was (except that Horn does not converge to anti-Horn, nor *vice versa*).

Of course, just by sheer definition, the Horn strategy could never dominate the anti-Horn in the simulations as described above; the meanings and the forms are not different in any way, and therefore both strategies are equivalent. However, by introducing conditions that follow from the definition of deviant and marked, the population can be made to converge towards the Horn strategy and not towards the anti-Horn strategy. These conditions are:

1. the use of the complex form is costly, in terms of points, and
2. the deviant meaning occurs less frequently than the normal meaning.

Thus the population is stimulated to use the simple form more frequently than the complex one. Since the normal meaning occurs more often than the deviant one it may be expected that the simple form will be linked to the normal meaning and the complex form to the deviant meaning: the characteristics of the Horn strategy. The conditions comply with Horn's description of "pragmatic labour": as little effort as possible will be made. However, this is less trivial than it seems. The use of the most economic strategy hardly makes any sense if one is not understood. Therefore the cost and benefits have to be balanced in some way. The assertion that frequency of the situations is non-identical, is not a complication of the original assumptions; it only formalises the distinction that

was already put forward in the original definition of Horn's division of pragmatic labour.

6.1 Adding costs to the model

The method to attribute points is somewhat extended:

$$\begin{aligned} \text{Points(a)} &= \text{Points(b)} = \\ &\sum_i P(i) [\delta(H_b(S_a(i)), i) - k(S_a(i)) + \delta(H_a(S_b(i)), i) - k(S_b(i))] \end{aligned}$$

$P(i)$ is the probability of meaning/situation i ; $k(f)$ gives a cost for signal f .

For the simple meaning we assume a probability of 1, the probability of the deviant meaning was varied between 0 and 1. In the same way the cost of the simple signal was assumed to be 0, and the cost of the complex signal between 0 and 1 (to prevent that a successful conversation would produce a negative yield).

6.2 Simulations with costs and probability added

Procedure

The procedure was the same as in the simulations of section 5, but with the cost function as just described. The cost differences are relatively low cost differences (0.8 for simple versus 1 for complex), as are the differences in the probability of normal and deviant meaning (0.8 for normal versus 1 for deviant). All four different simulations were repeated with cost and probability.

Results

When starting with the Smolensky strategy, the population eventually converges on the Horn strategy (although the number of evolutions needed for converges can be high when starting with the anti-Horn strategy). The anti-Horn strategy stops to be stable and a population of agents using the Horn strategy comes into being in 98% of the cases.

The simulations with mixed strategies end similarly. This also holds for start populations of Horn only (that do not change). It does not hold for the anti-Horn strategy under the present settings.⁴

⁴ It is possible to force the Horn-strategy to emerge from an anti-Horn population with a combination of extreme cost and probability differences and an extremely high mutation rate. This is disregarded, as it violates the assumptions of the modelling realistic evolution.

7 General conclusion

The simulations show that without cost the evolutionary system either converges towards the Horn strategy or towards the anti-Horn strategy. These are the only two strategies with which the users communicate in the best possible way; they are also Nash equilibriums.

A simple addition to our model of communicating agents, namely costs and probabilities, is enough to show why and how the division of pragmatic labour makes sense. The connection between improbable situations and more involved (costly) signals emerges from simple interactions between agents that individually do not decide on the division of pragmatic labour.

The simulations with cost show how the population can converge towards the Horn strategy. This strategy is more efficient than the anti-Horn strategy and costs less. It is a combination of the strategies “minimise cost” and “maximise utility”.

The results we obtained are not extremely surprising. It is easy to see that the Horn strategy yields most points under the given circumstances. However, the conclusion that evolution tends to go into the direction most favourable to the entire population is not trivial; whether the Horn strategy is indeed the optimal solution remains to be seen for every agent. In addition, as the prisoner’s dilemma shows, without prior conference a group of agents might not converge on the strategy that is optimal for the population.

The most interesting result is, however, that the evolution model is able to abandon an already chosen direction (a strategy used by the majority of the initial population) and to end up at the Horn strategy, for 15 of the 16 possible strategies. This all happens without explicit co-ordination by the agents; moreover, the agents themselves do not weigh the possibilities.

This research may be continued in a number of ways, none of which trivially lead to the same result, even though they are likely to show similar outcomes. A more interesting and realistic model is possible by having the agents to use more different signals and to put them in more different situations. In addition, it is more realistic if an agent would be able to choose from various signals, each with a certain probability of being sent in a given situation.

The simulations presented here assume that every agent speaks with the same frequency to any other agent. It would be logical to make agents speak to others located in their neighbourhood more than with agents far away. This would lead to subgroups that understand each other well, but members of different subgroups less well, as is the case in dialects.

For philosophy’s sake it would be interesting to research the value of testing a theory in this way, since the circumstances are dictated by the theory

itself. However, the theory used to be a principle that explained other semantic/pragmatic phenomena (like implicatures); the simulations and formalisations presented here show that the principle is consistent and does not need an assumption of innateness; it arises from very basic and assumedly uncontroversial formal translations of the definition itself. The non-iconic connection between simple form and normal meaning, paired with the coupling of marked form to marked meaning, emerges in evolution, and it does not have to be “designed in” anywhere in the individuals’ systems. It is both a possible outcome and the best outcome for signalling games in an evolutionary perspective.

Finally, it should be noted that similar results could be found by a paradigm called iterated learning (e.g. Kirby & Hurford, 2002) which can be seen as an alternative approach to cultural evolution. An important research objective is to adjust the existing methods of cultural evolution and to apply them to empirically investigated situations of language change. This necessitates first of all a clarification of the relationships between iterated learning and Steel’s recruitment theory (see Steels, 1998), as well as between the main internal constituents of either of them.

8 References

- Benz, A., Jäger, G., & van Rooij, R. (2005). *An introduction to game theory for linguists*. Houndsmills, Basingstoke, Hampshire: Palgrave Macmillan.
- Blutner, R., Borra, E., Lentz, T., Obdeijn, A., Uijlings, J. & Zevenhuijzen, R. (2002). Signalling Games: hoe evolutie optimale strategieën selecteert. In *Handelingen van de 24ste Nederlands-Vlaamse Filosofiedag*. Amsterdam: Universiteit van Amsterdam.
- Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University Of Chicago Press.
- Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11-42). Washington: Georgetown University Press.
- Kirby, S., & Hurford, J. (2002). The Emergence of Linguistic Structure: An overview of the Iterated Learning Model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the Evolution of Language* (pp. 121-148). London: Springer Verlag.
- Lewis, D. (1969). *Convention: A Philosophical Study*. Princeton: Harvard University Press.
- Nash, J. F. (1950). Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the United States of America* 36, 48-49.
- Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

- Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103, 133–156.
- Steels, L., & Belpaeme, T. (2004). Coordinating Perceptually Grounded Categories through Language. A Case Study for Colour. *Behavioral and Brain Sciences*.
- Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry* 29, 229-268.
- Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. Cambridge Mass.: MIT Press.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Mass.: Harvard University Press.
- Van Rooy, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy* 27, 493-527.
- Weibull, J. W. (1995). *Evolutionary Game Theory*. Cambridge, Mass.: MIT Press.

An Epistemic Interpretation of Bidirectional Optimality Based on Signaling Games

Michael Franke

Universiteit van Amsterdam

Institute for Logic, Language and Computation

Amsterdam, The Netherlands

To some, the relation between bidirectional optimality theory and game theory seems obvious: strong bidirectional optimality corresponds to Nash equilibrium in a strategic game (Dekker and van Rooij 2000). But in the domain of pragmatics this formally sound parallel is conceptually inadequate: the sequence of utterance and its interpretation cannot be modelled reasonably as a strategic game, because this would mean that speakers choose formulations independently of a meaning that they want to express, and that hearers choose an interpretation irrespective of an utterance that they have observed. Clearly, the sequence of utterance and interpretation requires a dynamic game model. One such model, and one that is widely studied and of manageable complexity, is a signaling game. This paper is therefore concerned with an epistemic interpretation of bidirectional optimality, both strong and weak, in terms of beliefs and strategies of players in a signaling game. In particular, I suggest that strong optimality may be regarded as a process of internal self-monitoring and that weak optimality corresponds to an iterated process of such self-monitoring. This latter process can be derived by assuming that agents act rationally to (possibly partial) beliefs in a self-monitoring opponent.

1 Bidirectional Optimality in Pragmatics

Optimality theory (or) has its origin in phonology (Prince and Smolensky 1997), but has been readily applied to other linguistic subdisciplines such as syntax, semantics (Hendriks and de Hoop 2001), and pragmatics (c.f. the contributions in Blutner and Zeevat 2004). Abstractly speaking, or is a model of how input and output representations are associated with each other based on grammatical preferences on input-output matching. More concretely, for models of prag-

matic interpretation we are interested in how a set M of (input) forms and a set T of (output) meanings are matched by language users in production and interpretation. An OT-SYSTEMS $\langle \text{Gen}, \geq \rangle$ is then just a pair $\langle \text{Gen}, \geq \rangle$ consisting of a GENERATOR $\text{Gen} \subseteq M \times T$ that gives us the initially possible form-meaning pairs and an ordering \geq on elements of Gen that measures how well the elements of the generator satisfy certain standards of grammaticality, normality, efficiency, or whatever might be at stake for a particular explanation of pragmatic language use.¹

Based on the ordering \geq , an OT-system specifies the preferred input-output associations in several ways. Since \geq is an ordering on a set of input-output pairs, we can either take a production perspective and ask which output is best when we fix the input dimension, or we can take a comprehension perspective and ask which input is best when we fix the output dimension. The former production perspective is taken by OT-syntax, the latter comprehension perspective is taken by OT-semantics. Abstractly, we can define the set of UNIDIRECTIONALLY OPTIMAL PAIRS as follows:

$$\begin{aligned} \text{OT}_{\text{syn}} &= \{ \langle m, t \rangle \in \text{Gen} \mid \neg \exists t' : \langle m, t' \rangle \in \text{Gen} \wedge \langle m, t' \rangle > \langle m, t \rangle \} \\ \text{OT}_{\text{sem}} &= \{ \langle m, t \rangle \in \text{Gen} \mid \neg \exists m' : \langle m', t \rangle \in \text{Gen} \wedge \langle m', t \rangle > \langle m, t \rangle \}. \end{aligned}$$

Optimization along both dimensions at the same time is also possible, of course. This is BIDIRECTIONAL OPTIMALITY and it comes in two varieties, a strong notion and a weak notion (Blutner 1998, 2000). We say that an input-output pair is STRONGLY OPTIMAL iff it is unidirectionally optimal for both production and comprehension:

$$\text{BIOT}_{\text{str}} = \text{OT}_{\text{syn}} \cap \text{OT}_{\text{sem}}$$

is the set of all strongly optimal pairs. Adopting Jäger's reformulation of Blutner's original definition (Jäger 2002), we say that a pair $\langle m, t \rangle$ is WEAKLY OPTIMAL iff

- (i) there is no weakly optimal $\langle m, t' \rangle$ such that $\langle m, t' \rangle > \langle m, t \rangle$; and
- (ii) there is no weakly optimal $\langle m', t \rangle$ such that $\langle m', t \rangle > \langle m, t \rangle$;

and we denote the set of all weakly optimal pairs with $\text{BIOT}_{\text{weak}}$. It is obvious

¹ Normally, the ordering \geq would be derived from a set of ranked constraints, but for the purposes of this paper we can safely abstract from that.

that all strongly optimal pairs are also weakly optimal, but it may be the case that there are weakly optimal pairs which are not strongly optimal.

How should we interpret the various optimality notions for applications to linguistic pragmatics? What exactly does it mean when an *or*-system selects a given form-meaning pair as weakly optimal but not strongly optimal, or as unidirectionally optimal but not strongly optimal? These are the general questions that this paper seeks to address.

Proponents of *or*-pragmatics are not unanimous about this issue. Some propose to think of unidirectional and strong optimality as measures of online pragmatic competence, but reject the notion that weak optimality has anything to do with actual pragmatic reasoning (Blutner and Zeevat 2004, 2008). Weak optimality is rather viewed from a diachronic, evolutionary perspective as giving the direction into which semantic meaning of expressions will most likely shift over time by pragmatic pressures.

Opposed to this view, others treat also weak optimality as a model of pragmatic reasoning competence. Under this interpretation different notions of optimality express different levels of *perspective taking*: whereas unidirectional optimization does not require to take the interlocutor's perspective into account, bidirectional optimization does (cf. Hendriks et al. 2007, chapter 5). More strongly even, optimality theory in pragmatics is often related to theory of mind (*toM*) reasoning (Premack and Woodruff 1978): unidirectional optimization is taken to involve no *toM* reasoning (or zero-order *toM*), strong optimization would correspond to first-order, and weak optimization would involve second-order *toM* reasoning (see, for instance, Flobbe et al. 2008, p. 424).

Given the controversy about its conceptual interpretation, what would be required is, in a manner of speaking, an interpretation of the basic notions of optimality theory that clarifies (some of) its intended use in pragmatic applications. For this purpose, it would be most welcome to supply in particular an *epistemic interpretation* of *or*, i.e., an interpretation that links *or*'s basic notions to more familiar features of human cognition such as beliefs and preferences. Comparison to a related game theoretic model can help achieve this, especially when we focus on an epistemic characterization of player behavior. This is what this paper tries to achieve by linking pragmatic *or*-systems to particular kinds of signaling games and by linking notions of optimality to particular player types of varying degree of sophistication.

The paper is structured as follows. I will first review critically the most commonly adopted characterization of or-pragmatics in terms of strategic games in section 2. It will transpire that a strategic game is inadequate to capture the sequential nature of speech and its uptake and interpretation. Section 3 explores a different characterization of or in terms of signaling games, and section XYZ finally links optimality notions to iterated best responses.

2 BiOT and Strategic Games

Bidirectional optimization is simultaneous optimization of both the production and the comprehension perspective. At first glance, this looks very similar to an equilibrium state in which the speaker's and the hearer's preferences are balanced. And, indeed, there is a *prima facie* very plausible link between BiOT and game theory. Dekker and van Rooij (2000) (henceforth D&vR) show that the notion of strong optimality corresponds one-to-one to the notion of Nash equilibrium in an *optimality game*.² An optimality game is a straightforward translation of an or-system into a strategic game. D&vR continue to show that weak optimality corresponds with the outcome of a process that we could call *iterated Nash-selection*. Let's first look at the analysis of D&vR in more detail and then reflect critically.

2.1 Strong Optimality as Nash Equilibrium

Formally a strategic game is a triple $\langle N, (A)_{i \in N}, (\succeq)_{i \in N} \rangle$ where N is a set of players, A_i are the actions available to player i and \succeq_i is player i 's preference relation over action profiles $\times_{j \in N} A_j$, i.e., possible outcomes of the game. A Nash equilibrium of a strategic game is an action profile a^* such that for all $i \in N$ there is no $a_i \in A_i$ for which:³

$$(a_{-i}^*, a_i) \succ_i a^*.$$

In words, a Nash equilibrium is an action profile which no player would like to deviate from given that all other players conform.

Take an or-system with forms M , meanings T —assuming for simplicity

² D&vR use the term “interpretation game” for what I call “optimality game.” I would like to reserve the former term for a particular kind of signaling game to be introduced later.

³ Here, (a_{-i}^*, a_i) is the action profile which is derived from a^* by replacing player i 's action in a^* with a_i .

that $\text{Gen} = M \times T$ — and some ordering \geq over form-meaning pairs. An **OPTIMALITY GAME**, as defined by D&vR, is a strategic game between a speaker S and a hearer H such that the speaker selects a form, $A_S = M$, the hearer selects a meaning, $A_H = T$, and the players' preferences are just equated with the ordering of the OT-system, $\geq_S = \geq_H = \geq$.

An action profile $\langle m, t \rangle$ is a Nash equilibrium of an optimality game iff

- (i) there is no $m' \in M$ such that $\langle m', t \rangle >_S \langle m, t \rangle$; and
- (ii) there is no $t' \in T$ such that $\langle m, t' \rangle >_H \langle m, t \rangle$.

But since $\geq_S = \geq_H = \geq$ this is the case just when $\langle m, t \rangle \in \text{BIOT}_{\text{str}}$. Consequently, every Nash equilibrium of an optimality game is a strongly optimal pair in the corresponding OT-system, and every strongly optimal pair of an OT-system is a Nash equilibrium of the corresponding optimality game. D&vR's result in slogan form: strong optimality is Nash equilibrium (in an optimality game).

2.2 Weak Optimality as Iterated Nash Selection

In order to understand D&vR's characterization of weak optimality, we should first notice that the recursive definition of weak optimality given above is rather cumbersome to apply. In practice, therefore, most often weakly optimal pairs are computed via a manageable algorithm which iteratively computes optimal pairs. D&vR's characterization of weak optimality is inspired by this iterative computation process, so that we should first revisit the **BIOT**-algorithm.

2.2.1 The **BIOT**-Algorithm

The **BIOT**-algorithm, which is due to Jäger (2002) and given in figure 1, iteratively computes three disjoint sets of form-meaning pairs (Jäger 2002):

- (i) the set Pool_n of form-meaning pairs still in competition for optimality after n rounds of iteration;
- (ii) the set Opt_n of form-meaning pairs that have been identified as optimal after round n ;
- (iii) the set Blo_n of form-meaning pairs that are blocked by an optimal pair and therefore removed from the pool.

```

Pool0 ← Gen
Opt0 ← ∅
Blo0 ← ∅
n ← 0
while Pooln ≠ ∅ do
  Optn+1 ← Optn ∪ {⟨m, t⟩ ∈ Pooln |
    ¬∃ ⟨m', t'⟩ ∈ Pooln ⟨m', t'⟩ > ⟨m, t⟩ ∧
    ¬∃ ⟨m, t'⟩ ∈ Pooln ⟨m, t'⟩ > ⟨m, t⟩}
  Blon+1 ← Blon ∪ {⟨m, t⟩ ∈ Pooln |
    ∃ ⟨m', t'⟩ ∈ Optn+1 ⟨m', t'⟩ > ⟨m, t⟩ ∨
    ∃ ⟨m, t'⟩ ∈ Optn+1 ⟨m, t'⟩ > ⟨m, t⟩}
  Pooln+1 ← Pool0 \ (Optn+1 ∪ Blon+1)
  n ← n + 1
end while

```

Figure 1: The BIOT-algorithm

Initially, Pool₀ is the set Gen and there are no optimal or blocked forms. The algorithm then iteratively computes optimal pairs based on a comparison of forms left in the pool and removes optimal and blocked pairs from the pool until every form-meaning pair is removed from the pool as either optimal or blocked. We could think of the pool at round n as a reduced OT-system. The BIOT-algorithm thus repeatedly checks for strong optimality in ever more reduced OT-systems and thus selects all and only weakly optimal pairs (see Jäger 2002; Franke 2009, for more formal detail).

2.2.2 Iterated Nash Selection

The main idea of D&vR's characterization of weak optimality is now this. Firstly, we saw that the BIOT-algorithm iteratively computes strongly optimal pairs, based on a shrinking pool of candidate pairs. Secondly, we also saw that strong optimality can be likened to Nash equilibrium in optimality games. Hence, the workings of the BIOT-algorithm can be recast in game theoretic terms as a process of iteratively removing action profiles from competition for Nash equilibrium that are, in a way of speaking, dominated by a Nash equilibrium.

In order to make this idea more precise, D&vR allow strategic games to have partial preferences. For games with partial preferences, not every definition of Nash equilibrium will do, but the one given above applies. The process of ITERATED NASH-SELECTION on a strategic game $I_0 = \langle N, (A)_{i \in N}, (\succeq_0)_{i \in N} \rangle$ is defined inductively as follows: let NE _{n} be the set of Nash equilibria of game I_n ;

I_{n+1} is derived from I_n by restricting the preferences $\succeq_{n,i}$ to:

$$\succeq_{n+1,i} = \{ \langle x, y \rangle \in \succeq_{n,i} \mid \neg \exists z \in \text{NE}_n : z \succ_{n,i} x \}.$$

If for some index n we have $I_n = I_{n+1}$, we consider the process to be terminated, and call NE_n the *outcome* of the process of iterated Nash-selection. D&vR show that this process corresponds to the BIOT-algorithm if applied to optimality games: if I is the optimality game corresponding to an OT-system, then the outcome of iterated Nash-selection on I contains all and only the weakly optimal pairs of the OT-system.

2.3 Critique

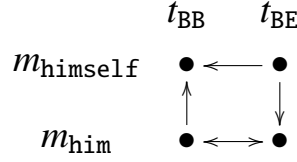
The characterization of strongly optimal pairs as Nash equilibria in an optimality game has some *prima facie* plausibility and seems unanimously endorsed as *the* link between OT and game theory. But on closer look the suggested parallel turns out not to be very sensible. To model communication as a strategic game is to assume that speakers choose formulations *independently* of a meaning that they want to express, and that hearers choose an interpretation *irrespective* of an utterance that they have observed. But this is clearly inadequate for pragmatic explanations. Obviously, speakers choose forms *conditional* on a meaning to be expressed and hearers choose interpretations of a *given* form not interpretations *per se*.

Here is a concrete example to make my argument more tangible. Let us consider the OT-system that Hendriks and Spence (2005) use in order to explain the preferred interpretations of sentences (1) and (2).

- (1) Bert washed himself.
- (2) Bert washed him.

Clearly, for (most) adult speakers of English the sentence (1) has only a coreferential reading for the reflexive pronoun, i.e., (1) means that Bert washed Bert. In contrast, sentence (2) has *no* coreferential reading for the non-reflexive pronoun, i.e., to (most) adult speakers (2) means that Bert washed someone other than himself. In order to model not only adult interpretation, but also a peculiar pattern in the acquisition of this piece of pragmatic competence, Hendriks and Spence (2005) adopt a simple OT-system with two forms m_{himself} for (1) and

m_{him} for (2), and two meanings t_{BB} for a situation in which Bert washed Bert and t_{BE} for a situation in which Bert washed Ernie. All possible form-meaning combinations are generated in this system and the ordering can be visualized as follows:⁴



This gives rise to the following sets of optimal pairs:

$$\begin{aligned}
 \text{Opt}_{\text{syn}} &= \{\langle m_{\text{himself}}, t_{\text{BB}} \rangle, \langle m_{\text{him}}, t_{\text{BE}} \rangle\} \\
 \text{Opt}_{\text{sem}} &= \{\langle m_{\text{himself}}, t_{\text{BB}} \rangle, \langle m_{\text{him}}, t_{\text{BB}} \rangle, \langle m_{\text{him}}, t_{\text{BE}} \rangle\} \\
 \text{BIOT}_{\text{str,weak}} &= \{\langle m_{\text{himself}}, t_{\text{BB}} \rangle, \langle m_{\text{him}}, t_{\text{BE}} \rangle\}
 \end{aligned}$$

We should now ask what it means to say that the strongly optimal pairs are $\langle m_{\text{himself}}, t_{\text{BB}} \rangle$ and $\langle m_{\text{him}}, t_{\text{BE}} \rangle$ and whether this squares with what it means to be a Nash equilibrium in an optimality game.

Suppose the speaker (Alice) and the hearer (Bob) are playing the corresponding strategic optimality game. In this game, both Alice and Bob make effectively simultaneous and *independent* decision. We may imagine that this is achieved, e.g., by writing down and passing to a judge the choice between either m_{him} or m_{himself} for Alice, and between t_{BB} and t_{BE} for Bob. A Nash equilibrium is then a pair of actions $\langle m, t \rangle$ such that, firstly, given Bob's choice t , Alice would not strictly prefer a message different from m , and, secondly, given Alice's choice m , Bob would not strictly prefer an interpretation different from t . This means that if this game is played repeatedly, and if, for example, Bob shows a tendency to play t_{BB} however slightly more frequently, then Alice would start to play m_{himself} more frequently, and the whole process would start reinforcing itself until we reach the *steady state* in which Alice always plays m_{himself} and Bob always plays t_{BB} . This steady state is a Nash equilibrium, and to think of Nash equilibria as steady states in this way is indeed the most prevalent textbook interpretation of this solution concept (e.g. Osborne and Ru-

⁴ An arrow from one form-meaning pair to another indicates that the form-meaning pair to which the arrow points is strictly more preferred according to \geq . It is not essential here that the ordering is derived from particular constraints, each with its own independent motivation (see Hendriks and Spenader 2005, for details).

binstein 1994; Osborne 2004; Heap and Varoufakis 2004).

But this has nothing to do with either the way that we imagine communication to proceed if online pragmatic reasoning is concerned, or with a reasonable model of language evolution under pragmatic pressures. It is also not the way we would commonly interpret a set of (strongly) optimal form-meaning pairs. The above set of strongly optimal pairs, which contains $\langle m_{\text{himself}}, t_{\text{BB}} \rangle$ and $\langle m_{\text{him}}, t_{\text{BE}} \rangle$, is commonly taken to describe *conditional* production and interpretation behavior (that is in a certain sense optimal). In particular, from a production point of view this set captures that if the speaker wants to express the meaning t_{BB} then it is optimal to use message m_{himself} and that if the speaker wants to express the meaning t_{BE} then it is optimal to use message m_{him} . Similarly, from an interpretation point of view the set captures that if the hearer observes message m_{himself} , he should optimally interpret this as meaning t_{BB} , and if the hearer observes message m_{him} , he should optimally interpret this as meaning t_{BE} .

In effect, that means that Nash equilibrium is an inadequate characterization of strong optimality, because optimality games are, qua *strategic* game, the inadequate game model for pragmatic OT-systems. The critique then carries over to Dekker and van Rooij’s characterization of weak optimality in terms of iterated Nash selection. Phrased polemically, if Nash equilibrium is an inadequate characterization of strong optimality, then if you repeatedly link Nash equilibrium and strong optimality in a reduced system (be it OT-system or optimality game), then this is not making things better, but worse.⁵

3 BiOT and Signaling Games

The above considerations suggest that the natural way of interpreting a set of form-meaning pairs —be they optimal or not— is not as a set of Nash equilibria, but rather as a (possibly partial) specification of *conditional* production and interpretation behavior. Speech production proceeds from a thought or intention that needs to be expressed to a choice of form to express the desired content with. Interpretation of an utterance starts only after a message that needs to be interpreted has been observed. This is all natural, I believe, but it does call for a different game model to match pragmatic OT-systems: we need at least a

⁵ For more detailed criticism also of the concept of iterated Nash selection see Franke (2009).

dynamic game in which the speaker chooses a message conditional on a to-be-expressed meaning, and the hearer subsequently chooses an interpretation given that he has observed a form.

The perhaps most manageable and (for that reason) most widely studied kind of game that fits this description is a *signaling game*. A signaling game is a special kind of dynamic game with incomplete information that has been studied extensively in philosophy (Lewis 1969), economics (Spence 1973), biology (Zahavi 1975; Grafen 1990) and linguistics (Parikh 1991, 1992, 2001; van Rooij 2004). Informally speaking, the idea is that the sender (the agent modelling the speaker) knows the true state of affairs t , but the receiver (the agent modelling the hearer) does not. Given the true state t the sender then chooses a message m which the receiver observes. Subsequently, the receiver chooses an action a as his proper response. An outcome of such a game is given as the triple $\langle t, m, a \rangle$. Naturally, sender and receiver may prefer some outcomes more than others and these preferences may select for a particular class of sender and receiver behavior under a given solution concept.

Formally, a signaling game (with meaningful signals) is a tuple

$$\langle \{S, R\}, T, \text{Pr}, M, \llbracket \cdot \rrbracket, A, U_S, U_R \rangle$$

where sender S and receiver R are the players of the game; T is a set of states of the world; $\text{Pr} \in \Delta(T)$ is a probability distribution over T , which represents the receiver's uncertainty which state in T is actual;⁶ M is a set of messages that the sender can send; $\llbracket \cdot \rrbracket : M \rightarrow \mathcal{P}(T) \setminus \emptyset$ is a denotation function that gives the predefined semantic meaning of a message as the set of all states where that message is true (or otherwise semantically acceptable); A is the set of response actions available to the receiver; and $U_{S,R} : T \times M \times A \rightarrow \mathbb{R}$ are utility functions for both sender and receiver that give a numerical value for, roughly, the desirability of each possible play of the game.⁷

In general, behavior of players in dynamic games is represented in terms

⁶ As for notation, $\Delta(X)$ is the set of all probability distributions over set X , Y^X is the set of all functions from X to Y , $X : Y \rightarrow Z$ is alternative notation for $X \in Z^Y$, and $\mathcal{P}(X)$ is the power set of X .

⁷ To rule out certain irrelevant and aberrant cases, I will assume throughout that for each state t there is at least one message m such that $t \in \llbracket m \rrbracket$ and that Pr has full support, i.e., that $\text{Pr}(t) > 0$ for all $t \in T$.

of STRATEGIES which select possible moves for each agent for any of their choice points in the game. For signaling games, a PURE SENDER STRATEGY $s \in M^T$ is a function from states to messages which specifies which message the sender will or would send in each state that might become actual. A PURE RECEIVER STRATEGY $r \in A^M$ is a function from messages to actions which similarly specifies which action the receiver will or would choose as a response to each message he might observe. (Obviously, the receiver knows only what message has been sent, but not what state is actual, so he has to choose an action for each message he might observe and cannot condition his choice on the actual state of affairs). A PURE STRATEGY PROFILE $\langle s, r \rangle$ is then a characterization of the players' *joint behavior* in a given signaling game.

3.1 Optimal Pairs as Partial Strategies

If my previous argument is correct, and a set of optimal pairs, is to be interpreted as a specification of *conditional* production or comprehension behavior, then we should generally link sets of form-meaning pairs, be they optimal or not, to strategies in a suitable signaling game. In particular, a set of form-meaning pairs partially defines a sender or receiver strategy in a SIGNALING GAME WITH INTERPRETATION ACTIONS where

- (i) the set of states in the signaling game are the meanings T of the OT-system; these are the meanings that the speaker might want to express;
- (ii) the set of messages in the signaling game are the forms M of the OT-system; these are the messages the speaker can choose to express a meaning when she wants to; and
- (iii) the set of receiver actions in the signaling game are interpretations, i.e., the meanings T of the OT-system.

In general, we can read off a (partial) description of a sender and receiver strategy for such a game from any set $O \subseteq M \times T$. If we agree to write

$$O(t) = \{m \in M \mid \langle m, t \rangle \in O\} \quad \text{and} \quad O(m) = \{t \in T \mid \langle m, t \rangle \in O\}, \quad (3.1)$$

the set of pure sender strategies in a signaling game with interpretation actions compatible with O is:

$$S(O) = \{s \in \mathbf{S} \mid O(t) \neq \emptyset \rightarrow s(t) \in O(t)\};$$

and the set of pure receiver strategies compatible with O is:

$$R(O) = \{r \in \mathbf{R} \mid O(m) \neq \emptyset \rightarrow r(m) \in O(m)\}.$$

Obviously, an arbitrary set O need not specify a full strategy. For instance, there may be states t for which $O(t)$ is empty, so that when taken as a description of a sender strategy O is only a *partial* description. I suggest that this is really how we should set the link between or and game theory in pragmatics: sets of form-meaning pairs —no matter whether any notion of optimality has selected these— are specifications of strategies in a corresponding signaling game with interpretation actions.

3.2 OT-Systems and Signaling Games

Linking form-meaning pairs to strategies may be a natural idea, but this much does not yet fix a complete translation between or-systems and signaling games. Some correspondences are hardly worth mentioning: speakers correspond to senders and hearers correspond to receivers, of course, and the generator places restrictions on the set of possible form-meaning associations and this naturally finds its expression in the semantic denotation function

$$\langle m, t \rangle \in \text{Gen} \text{ iff } t \in \llbracket m \rrbracket$$

if we assume that the corresponding signaling game makes truthful signaling obligatory, i.e., that the sender can only ever send a true message in a given state. But all this still does not fix an interpretation of the ordering \geq of the or-system. Also the prior probabilities $\text{Pr}(\cdot)$ and the utilities $U_{S,R}$ for both sender and receiver are still unspecified.

Formally, there are many possibilities of translation between or-systems and signaling games. I have explored one such formal parallel in Franke (2009),

where I link optimality notions with the behavior of strategic types in a sequence of iterated best responses. An iterated best response model, or *IBR* model for short, is an epistemic solution concept in which different strategic types of players are defined in terms of their beliefs about opponent player behavior (cf. Jäger 2008; Jäger and Ebert 2009). The beginning of the sequence is given by naïve strategic types of level 0 who do not take their opponent's perspective into account, but who may be susceptible to certain focal framing effects in the game structure (cf. Schelling 1960, for focality). Players of level $k+1$ then believe that they are facing a level- k opponent and play a best response to that belief. It is then possible to identify in particular the behavior of naïve level-0 receivers with unidirectionally optimal interpretation, the behavior of level-1 senders with unidirectionally optimal production and the interpretation of level-2 receivers with strong optimality if we assume that the receiver uses a particular, simplistic (and strictly speaking incorrect) belief formation process when computing his posterior beliefs after receiving a message (see Franke 2009, for details).

This characterization of optimality notions in terms of *IBR* reasoning assumed an independently motivated *IBR* model and tried to match optimality notions with as little amendment as possible onto the strategic types of this model. It turned out, however, that especially a characterization of weak optimality is rather difficult, because the game-theoretic idea of a rational best response to a belief in an opponent strategy is *holistic* in the sense that it takes into account the whole of an opponents strategy (see also section 4.2). This makes it possible that certain form-meaning associations appear optimal in early stages, but are dismissed as optimal later on, because every possible form-meaning association is always reconsidered at every iteration step. Opposed to that, the *BIOT* algorithm, which selects for weak optimality, is rather *myopic* in that the set of optimal form-meaning associations grows monotonically. The upshot of this is that Bayesian rationality, if based on a standard belief in opponent strategy, does not always match the fast-and-frugal form-meaning selection process modelled by the *BIOT* algorithm. In Franke (2009) I therefore give a restriction on agent's belief formation, which is admittedly rather severe, but which guarantees a match between rationalistic *IBR* and weak optimality.

In the following, I would like to go a different route, one that is closer to *OT*-pragmatics and parts from the idea of staying as close as possible to the rationalistic norms of standard game theory. I would like to start out from the

assumption that sets of form-meaning pairs describe partial strategies of senders and receivers in a signaling game. Based on this, it is possible to simply reconstruct the BIOT-algorithm as a *behavioral definition* of strategic types of players. Finally, we can then look back at this behavioral characterization and ask which *epistemic assumptions*, e.g., about belief formation, rationality or preferences, would give rise to this behavior and how these assumptions square with the common interpretation of optimality notions in the pragmatic OT-community on the one hand, and the accepted standards of game theory on the other. The epistemic interpretation of optimality that I end up suggesting is that bidirectional optimality is a process of, if necessary iterative, self-monitoring for congruence between form-meaning associations in production and interpretation.

4 Iterated Self-Monitoring

In order to match sets of form-meaning pairs to strategies of senders and receivers, we should assume that the set of receiver actions equals the set of states $T = A$. Going a step further, let us also assume that the signaling game corresponding to a given OT-system has a particular payoff structure, namely that the signaling game models a situation in which sender and receiver would like to communicate the true state of affairs successfully. This is achieved by setting:

$$U_S(t, m, a) = U_R(t, m, a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{otherwise.} \end{cases}$$

Let us call a signaling game with this payoff structure an INTERPRETATION GAME.

Recall that according to the standard interpretation of unidirectional optimality, as outlined in section 1, we want to link unidirectional optimality to the behavior of senders and receivers who do *not* take their opponent's strategy into account but only follow their own preferences on form-meaning associations as specified by a given OT-system. This can be achieved if we assume that there are *naïve strategic types* which do not take a belief about their opponent into account, but merely play a rational best response given their preferences about form-meaning associations.

4.1 Unidirectional Optimality and Naïve Players

For the sender this is easily achieved by assuming that messages have *state-dependent costs*. We model this by a function $C : T \times M \rightarrow \mathbb{R}$ that associates for every state t and message m the costs $C(t, m)$ that sending m in state t incurs for the sender.⁸ To translate the speaker's preferences, as captured in \geq into the signaling game, we simply assume that for all $\langle m, t \rangle$ and $\langle m', t \rangle$ in Gen:

$$\langle m, t \rangle \geq \langle m', t \rangle \text{ iff } C(t, m) \leq C(t, m').$$

We may then assume that a naïve, but rational sender type S_0 , who does not take interpretation behavior into account but otherwise cares for her preferences, will choose a message that minimizes costs in each state. We can represent this sender type by a set of pure strategies as follows:

$$S_0 = \left\{ s \in \mathbf{S} \mid \forall t \in T \ s(t) \in \arg \min_{m \in M} C(t, m) \right\}.$$

By construction, it is trivially so that:⁹

$$\langle m, t \rangle \in \text{OT}_{\text{syn}} \text{ iff } m \in S_0(t).$$

In words, our naïve sender type corresponds behaviorally to unidirectional optimality along the production dimension.

For the receiver a similar move is possible. Since in an interpretation game states correspond one-to-one to actions, and, moreover, the receiver would like to match his response action to the true state, we find that a receiver who does not take his opponent's strategy into account would maximize for the most likely state in which a given message could have been sent (given the restrictions on truthful signaling). That is to say that *prima facie* we would like to construct a naïve receiver type similar to S_0 who takes into account only his preferences as represented in his prior probabilities $\text{Pr}(\cdot)$. The problem with this is that not all OT-orderings can be translated in this way because, obviously,

⁸ I will follow standard practice and assume that these costs are nominal, i.e., that they apply only when expected utilities based on U_S reach a tie.

⁹ As for notation, a set of pure sender strategies like S_0 can equivalently be represented as a set of form-meaning pairs. With this, $S_0(t)$ is defined by the notational convention in (3.1).

$\text{Pr}(\cdot)$ only specifies a *global* ordering on T independent of the message that the receiver observes. However, there are independent arguments for thinking of the receiver's prior probabilities merely as a simplistic and convenient way of specifying those global form-meaning associations that do not vary with the message (see Franke 2009, section 3.1). If we then want to be able to translate any arbitrary σ -ordering into a signaling game via prior probabilities, we should adapt the definition of a signaling game to include an ASSOCIATION FUNCTION $\text{Ass} : M \times T \rightarrow \mathbb{R}$, such that for all $\langle m, t \rangle$ and $\langle m, t' \rangle$ in Gen we have:¹⁰

$$\langle m, t \rangle \geq \langle m, t' \rangle \text{ iff } \text{Ass}(m, t) \geq \text{Ass}(m, t').$$

Based on his associative preferences, a naïve receiver R_0 , who is rational but does not take into account his opponent's behavior, will maximize for each observed message the likelihood of matching the true state by selecting a maximally associated state:

$$R_0 = \left\{ r \in \mathbb{R} \mid \forall m \in M \ r(m) \in \arg \max_{t \in T} \text{Ass}(m, t) \right\}.$$

Again, by construction, this corresponds with unidirectional optimality along the comprehension dimension:

$$\langle m, t \rangle \in \sigma_{\text{sem}} \text{ iff } t \in R_0(m).$$

In line with the common idea that unidirectional optimization does not involve taking the opponent's behavior into account, the above definition of naïve players offers a straightforward behavioral implementation of unidirectional optimality in a signaling game. Moreover, this characterization also allows to draw further conclusions about a possible epistemic interpretation of unidirectional optimality. We should think of preferences, as captured in the σ -ordering \geq , as the strength of associating form-meaning pairs. This is given by grammar, in a wide sense of the term, and may involve contextual association biases, depending on the intended application of the σ -system. But, crucially, building on these basic grammatical preferences, unidirectional optimality is supplied by

¹⁰ A prior probability function $\text{Pr}(\cdot)$ is then just a special case of an association function: constant over all m and scaled to the interval $[0; 1]$.

Bayesian rationality in the absence of any conjecture about opponent behavior.

4.2 Strong Optimality as Self-Monitoring

Strong optimality is defined as the intersection of unidirectional optimization along the comprehension and the production dimension, and is therefore often considered an operation that takes into account the opponent's strategy (e.g. Hendriks and Spengler 2005; Flobbe et al. 2008). However, there is a fundamental difference between the way game theory models such perspective taking from the way this notion is present in strong and weak optimality. This section therefore suggests to look at strong optimality as a mere self-monitoring, not as genuine perspective taking in the strong game-theoretic sense.

Let us begin by looking more closely at the idea of perspective taking in game theory. If a rational agent takes the behavior of an opponent into account, game theorists assume that the agent plays a rational best response to the belief that her opponent is behaving in the specified way. Take, for instance, the behavior of a naïve receiver R_0 . A belief in this behavior is a belief that message m is interpreted as some state in $R_0(m)$. If a sender plays a best response to this belief, she optimizes her behavior, based on her preferences, by taking into account the complete interpretation behavior of R_0 , i.e., the way *all* messages are interpreted according to R_0 . In other words, perspective taking in game theory is *holistic* in the sense that the *whole* strategy of the opponent is taken into account when making a choice.

This is not what strong optimality implements. In order to adhere to strong optimality, it is usually not necessary to take the whole strategy of the opponent into account. For example, a sender only has to do two things if she wants to conform to strong optimality (when this is possible): firstly, given a state t , she needs to check her production preferences to compute $OT_{\text{syn}}(t) \subseteq M$; secondly, she has to check whether some message in $OT_{\text{syn}}(t)$ would also be interpreted as t given the receiver's interpretative preferences. It becomes clear thus that strong optimization merely implements a simple *associative feedback-loop*, but not full perspective-taking in the standard game-theoretic sense. In other words, under this interpretation strong optimality is mere SELF-MONITORING to check for *association congruence* between production and comprehension.

4.3 Weak Optimality as Iterated Self-Monitoring

This idea of monitoring production by self-interpretation and monitoring interpretation by self-production also carries over to an interpretation of weak optimality. Remember that the BIOT-algorithm repeatedly checks for strong optimality in reduced OT-systems where optimal and blocked form-meaning pairs are removed in every step. This process can be mirrored by defining more sophisticated player types of level $n > 0$ whose behavior corresponds to the n -th round of computation of the BIOT-algorithm.

For this purpose, let AC_0 be the set of *level-0 association congruent from-meaning pairs*: $AC_0 = \text{BIOT}_{\text{str}}$. Let us then define level- $(n + 1)$ players as playing in conformity with level- n association congruence where possible:¹¹

$$S_{n+1}(t) = \begin{cases} AC_n(t) & \text{if } AC_n(t) \neq \emptyset \\ \arg \min_{m \in M \setminus AC_n(T)} C(t, m) & \text{otherwise} \end{cases}$$

$$R_{n+1}(m) = \begin{cases} AC_n(m) & \text{if } AC_n(m) \neq \emptyset \\ \arg \max_{t \in T \setminus AC_n(M)} \text{Ass}(m, t) & \text{otherwise.} \end{cases}$$

To complete the inductive construction, we also need to define *level- $(n + 1)$ association congruence* as: $AC_{n+1} = S_n \cap R_n$. This construction, call it *iterated self-monitoring*, quite obviously replicates exactly the workings of the BIOT-algorithm.

Iterated self-monitoring is not only a rephrasing of the BIOT-algorithm, but actually helps interpreting weak optimality. For we can now ask and answer the question which assumptions about the psychology of agents give rise to the above behavior of sophisticated players. The obvious answer is that agents perform self-monitoring iteratively, but only when necessary, and believe that their opponents do too. More concretely, the behavior of a sophisticated level- $(n + 1)$ sender follows from two simple assumptions:

- (i) the player performs self-monitoring based on the behavior of level- n players and plays accordingly when this gives a result;
- (ii) where this gives no result, the player plays rationally given the *partial*

¹¹ I write $AC_n(T)$ as the set of all m for which there is some t such that $\langle m, t \rangle \in AC_n$, and similarly for $AC_n(M)$.

belief that the opponent adheres to (i).

Let us first validate that these two assumptions indeed give rise to the behavior of sophisticated players as defined above and reflect on the conceptual implications afterwards.

Take, for instance, a sender of level $(n + 1)$ who wishes to express the state t . (The argument for the receiver is parallel.) Firstly, S_{n+1} would perform self-monitoring based on level- n behavior and thus compute level- n association congruence. If some message satisfies level- n association congruence for t , any message with this property would be used. This way, the first assumption directly assures that $S_{n+1}(t) = AC_n(t)$ whenever $AC_n(t) \neq \emptyset$.

The second assumption is just a little bit more complicated. It kicks in when the sender wants to express some $t \notin AC_n(M)$. In that case, S_{n+1} is required to play rationally to the belief that her opponent's behavior is characterized by the (possibly partial) strategy $R_{n+1}(m) = AC_n(m)$ for all m such that $AC_n(m) \neq \emptyset$.¹² Given such a partial conjecture, it would always yield an expected utility of 0 (possibly minus some nominal cost, of course) to try to express a state in $t \notin AC_n(M)$ with a message $m \in AC_n(T)$. But in the absence of a definite conjecture about how messages in $M \setminus AC_n(T)$ are interpreted, any such message has at least a positive chance of obtaining the right interpretation, so that the expected utility of sending a message from the $M \setminus AC_n(T)$ in t will be strictly bigger than zero, and, in fact, equal for all messages in this set. Consequently, a rational level- $(n + 1)$ sender will choose any cost-minimal message in $M \setminus AC_n(T)$ in each state $t \notin AC_0(M)$. It turns out that the second assumption effectively gives, via partiality of belief in self-monitoring, a rationalistic explanation of the blocking mechanism of the BIOT-algorithm.

4.4 Reflection on Iterated Self-Monitoring

Taken together, this suggests that we should think of weak optimality as a process of self-monitoring to the maximal depth necessary to express or interpret a form. Since $AC_n \subseteq AC_{n+1}$ for all n , there is no need to compute more sophisticated play than the minimal k for which $AC_k(t) \neq \emptyset$, when expressing t , or $AC_k(m) \neq \emptyset$, when interpreting m . Only when necessary, further iteration

¹² Notice that this belief may be *partial*, for it may mean that S_{n+1} has no belief about how her opponent will interpret a message m for which $AC_n(m) = \emptyset$ if such messages exist.

of self-monitoring takes place, by adopting a belief that the opponent also performs such iterated self-monitoring. At each step of this procedure, however, the conjecture about opponent behavior is not the full-fledged perspective taking that is standard in game theory, but only an associative feedback and the assumption that the opponent also performs such self-monitoring.

Interestingly enough, Bayesian rationality features in this interpretation of optimality only as an explication of preference maximization *in the absence* of a conjecture about opponent behavior. In other words, unlike in the structurally similar IBR models of, for instance, Jäger and Ebert (2009) and Franke (2009), the more sophisticated types do not rely on deeper and deeper nestings of belief in rationality. The sophisticated types that match the BIOT-algorithm only require ever more nested beliefs in self-monitoring. This is in a sense a weaker requirement, but it may nonetheless explain why weak optimality is often too strong a theoretical prediction to be borne out in reality (cf. the arguments by Beaver and Lee 2004): that agents can coordinate successfully on weakly optimal communication behavior becomes dubious proportional to the number of iteration steps in self-monitoring, due to natural restrictions on cognitive resources.

However, to say that nested belief in self-monitoring, as found in BIOT under the interpretation favored here, is weaker than nested belief in rationality under full-fledged perspective taking, as found in recent IBR models, is not necessarily an argument *for* BIOT and against IBR. In order to be an argument for BIOT we would have to motivate why exactly this kind of self-monitoring should occur in pragmatic language use. It is fairly standard to assume monitoring by internal self-interpretation (cf. Levelt 1989), but this is not necessarily so for comprehension. This points favorably into the direction of an asymmetric approach to BIOT, as advanced by (see Zeevat 2000).

Finally, it is also not implausible to accept simple self-monitoring as a reasonable mental operation, be it in production alone or also in interpretation, yet to reject nested beliefs in self-monitoring opponents as a natural cognitive process. This would corroborate the position of, for instance, Blutner and Zeevat (2008) that only strong optimality is reasonable as an online mechanism, while weak optimality is not.

5 Conclusion

To take stock, I have argued that it useful and desirable to match optimality theory with game theory in order to supply a characterization of *or*'s basic notions in terms of agents' mental states and behavioral disposition. I have tried to show that the analogy between *or*-systems and strategic optimality games suggested by Dekker and van Rooij (2000) is conceptually flawed, and does not achieve this end. Therefore, I have suggested to work out a connection between signaling games and *or*-systems, and between optimality notions and different kinds of more or less sophisticated player types. From this point of view, unidirectional optimality is Bayesian rationality that takes into account only preferences on form-meaning associations in the absence of a conjecture about opponent behavior. Strong optimality turned out to be best described as a simple self-monitoring feedback process, not as full strategic perspective taking. Weak optimality then presents itself as an iterated process of such self-monitoring which is defined in terms of beliefs in self-monitoring and rational responses to these (possibly partial) beliefs.

References

- Beaver, David and Hanjung Lee (2004). "Input-Output Mismatches in Optimality Theory". In: *Optimality Theory and Pragmatics*. Ed. by Reinhard Blutner and Henk Zeevat. Palgrave MacMillan. Chap. 6, pp. 112–153.
- Blutner, Reinhard (1998). "Lexical Pragmatics". In: *Journal of Semantics* 15. Pp. 115–162.
- (2000). "Some Aspects of Optimality in Natural Language Interpretation". In: *Journal of Semantics* 17. Pp. 189–216.
- Blutner, Reinhard and Henk Zeevat, eds. (2004). *Optimality Theory and Pragmatics*. Palgrave MacMillan.
- (2008). "Optimality-Theoretic Pragmatics". To appear in: Claudia Maienborn, Klaus von Stechow and Paul Portner (eds.) *Semantics: An International Handbook of Natural Language Meaning*.
- Dekker, Paul and Robert van Rooij (2000). "Bi-Directional Optimality Theory: An Application of Game Theory". In: *Journal of Semantics* 17. Pp. 217–242.
- Flobbe, Liesbeth et al. (2008). "Children's Application of Theory of Mind in Reasoning and Language". In: *Journal of Logic, Language and Information* 17. Pp. 417–442.

- Franke, Michael (2009). “Signal to Act: Game Theory in Pragmatics”. PhD thesis. Universiteit van Amsterdam.
- Grafen, Alan (1990). “Biological Signals as Handicaps”. In: *Journal of Theoretical Biology* 144. Pp. 517–546.
- Heap, Shaun P. Hargreaves and Yanis Varoufakis (2004). *Game Theory — A Critical Text (Second Edition)*. Routledge.
- Hendriks, Petra and Helen de Hoop (2001). “Optimality Theoretic Semantics”. In: *Linguistics and Philosophy* 24. Pp. 1–32.
- Hendriks, Petra and Jennifer Spenader (2005). “When Production Precedes Comprehension: An Optimization Approach to the Acquisition of Pronouns”. In: *Language Acquisition* 13.4. Pp. 319–348.
- Hendriks, Petra et al. (2007). “Conflicts in Interpretation”. Unpublished book manuscript, Groningen, Nijmegen, Utrecht.
- Jäger, Gerhard (2002). “Some Notes on the Formal Properties of Bidirectional Optimality Theory”. In: *Journal of Logic, Language and Information* 11.4. Pp. 427–451.
- (2008). “Game Theory in Semantics and Pragmatics”. Unpublished manuscript, University of Bielefeld.
- Jäger, Gerhard and Christian Ebert (2009). “Pragmatic Rationalizability”. In: *Proceedings of Sinn und Bedeutung* 13. Ed. by Arndt Riester and Torgim Solstad. Pp. 1–15.
- Levelt, Willem J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press.
- Lewis, David (1969). *Convention. A Philosophical Study*. Harvard University Press.
- Osborne, Martin J. (2004). *An Introduction to Game Theory*. New York: Oxford University Press.
- Osborne, Martin J. and Ariel Rubinstein (1994). *A Course in Game Theory*. MIT Press.
- Parikh, Prashant (1991). “Communication and Strategic Inference”. In: *Linguistics and Philosophy* 14. Pp. 473–514.
- (1992). “A Game-Theoretic Account of Implicature”. In: *TARK '92: Proceedings of the 4th conference on Theoretical aspects of reasoning about knowledge*. San Francisco: Morgan Kaufmann Publishers Inc. Pp. 85–94.
- (2001). *The Use of Language*. Stanford University: CSLI Publications.
- Premack, David and Guy Woodruff (1978). “Does the Chimpanzee have a Theory of Mind”. In: *Behavioral and Brain Sciences* 1.4. Pp. 515–526.

- Prince, Alan and Paul Smolensky (1997). “Optimality: From Neural Networks to Universal Grammar”. In: *Science* 275. Pp. 1604–1610.
- van Rooij, Robert (2004). “Signalling Games Select Horn-Strategies”. In: *Linguistics and Philosophy* 27. Pp. 493–527.
- Schelling, Thomas C. (1960). *The Strategy of Conflict*. Cambridge, Massachusetts: Harvard University Press.
- Spence, Andrew Michael (1973). “Job market signaling”. In: *Quarterly Journal of Economics* 87. Pp. 355–374.
- Zahavi, Amotz (1975). “Mate Selection — A Selection for a Handicap”. In: *Journal of Theoretical Biology* 53. Pp. 205–214.
- Zeevat, Henk (2000). “The Asymmetry of Optimality Theoretic Syntax and Semantics”. In: *Journal of Semantics* 17.3. Pp. 243–262.

History and Grammaticalisation of "Doch"/"Toch"

Henk Zeevat

ILLC, University of Amsterdam

Elena Karagjosova

University of Stuttgart

The paper investigates the origins of the German/Dutch particle *toch/doch* in the hope of shedding light on a puzzle with respect to *doch/toch* and to shed some light on two theoretical issues. The puzzle is the nearly opposite meaning of the stressed and unstressed versions of the particle which cannot be accounted for in standard theories of the meaning of stress. One theoretical issue concerns the meaning of stress: whether it is possible to reduce the semantic contribution of a stressed item to the meaning of the item and the meaning of stress. The second issue is whether the complex use of a particle like *doch/toch* can be seen as an instance of spread or whether it has to be seen as having a core meaning which is differentiated by pragmatics operating in different contexts.

We use the etymology of *doch* and *toch* as *to+u+h* (that+ question marker+ emphatic marker) to argue for an origin as a question tag checking a hearer opinion. Stress on the tag indicates an opposite opinion (of the common ground or the speaker) and this sets apart two groups of uses spreading in different directions. This solves the puzzle, indicates that the assumption of spread is useful and offers a subtle correction of the interpretation of stress. While stress always means contrast with a contrasting item, if the particle use is due to spread, it is not guaranteed that the unstressed particle has a corresponding use (or inversely).

1 Introduction

Dutch *toch* or German *doch* give rise to an almost paradoxical question, first noted by Doherty(1985). If the sentence is presented with stress on *toch/doch* the conditions of use become the opposite from the same sentences with the stress removed.

- (1) Hij komt TOCH.
Er kommt DOCH.

He is coming, although we believed he was not.

- (2) Hij komt toch.
Er kommt doch.
You know he's coming.

The first speech act (1) is a correction, normally of the common ground between speaker and hearer and the second speech act (2) is a reminder of some common ground fact. The problem is how to derive these two different uses from the same conceptual source (the meaning of *doch/toch*) and a general account of the import of accent in interpretation. It is not an easy problem, since in Rooth's account of accent (Rooth 1992) all that accent contributes is the salience of an alternative to the accented part, here the particle. That works fine in (1), since the interlocutors believed (3), a clear alternative in the sense of Rooth. It can therefore be explained why accent appears on the particle in (1) but the explanation fails completely to predict why a quite distinct speech act results when the accent is omitted. One should be back at the core meaning without a salient alternative. But what is a correction of the common ground without a salient element of the common ground to be corrected? And what would it mean to remind somebody of a known fact while making its negation salient?

- (3) Hij komt NIET.
Er kommt NICHT.
He is NOT coming.

Notice that in (2) other alternatives can be salient.

- (4) PETER war doch in Frankfurt.
You know it was Peter who was in Frankfurt.

(4) will have a salient alternative, say *You know it was John who was in Frankfurt*, so the problem is not that reminders cannot have salient alternatives. The problem is that (3) is the only good alternative to (2) and assuming it is salient seems to destroy the point of the reminder and moreover does not lead to the meaning of (1). It would appear that the particle with accent and the particle without accent have acquired independent meanings and that -contrary to what is generally assumed- Dutch and German are distinguishing words by means of word intonation alone.

In this paper, we explore the history of the particle in order to solve this problem, largely in order to see whether progress can be made by applying the ideas in Zeevat (2007) about *spread* in language evolution. Our claim is that in

its original meaning the intonational contrast is as expected. Spread of uses of both the accented and unaccented *toch/doch* has resulted in a set of new uses, related to but distinct from the original use. Certain of those uses cause accent to appear because there is a contrasting and activated alternative, while other uses rule out accent. This results in accent (with the syntactic position of the particle) being one of the factors that disambiguates between the different uses of *doch/toch*. Accent in this view does not have its own meaning, but it has a triggering condition. The triggering condition is compatible with only some of the possible uses. There could be no compositional theory that takes the meaning of *doch* and the meaning of accent and combines them into the meaning of accented *doch*.

In section (2), an overview is presented of the different German and Dutch uses. Section (3) is about the view that words like *toch/doch* can be described by means of a core meaning. Section (4) discusses the alternative model of spread in the evolution of languages. Section (5) provides a possible historical explanation.

2 Overview of the uses of doch and toch

This section is an overview of the uses of *doch* and *toch* in German and Dutch. While there is a large overlap, there are also differences. Labels are introduced for the uses and these labels are used later on in the text. The overview is close to Foolen (2003), but adds some differences between Dutch and German.

Questions with assertion syntax

Asking for confirmation of something the other speaker said, prompted by having the opposite information (**correction confirmation question**):

- (5) Hij komt DOCH?
Er kommt DOCH?
Is he coming after all?

Confirmation of old common ground information, to make sure, or because the other seems to have forgotten (**reminder question**):

- (6) Hij komt toch?
Er kommt doch?
You know he's coming, isn't he?

- (7) Ik ga toch 2 weken weg?

Ich bin doch weg die naechste zwei Wochen?
You know I am away the next two weeks?

It is one of the most remarkable properties of *doch/toch* that it can make assertive sentences into questions. Unaccented *doch/toch* cannot be used with inversion, at least in polar questions.

Questions with question syntax

Asking for confirmation of correction of common ground: you were not coming but now you seem to be. They are not different from the corresponding question without inversion:

- (8) Kom je TOCH ?
Kommst du (also) DOCH ?
So you are coming after all?

Inversion is not possible with unaccented *toch/doch*:

- (9) *Kom je toch?
*Kommst du (also) doch?

Assertions

Correction of common ground:

- (10) Hij komt TOCH.
Er kommt DOCH.
He is coming after all.

Correction particle

TOCH WEL and DOCH: correction of negation:

- (11) TOch WEL/NIET)!
DOCH (NICHT)!
No!

Proconcessive use:

The previous context contains a reason for thinking otherwise, which must be

identified in a proper interpretation. The accent is weaker than in the correction cases.

- (12) En TOCH kwam hij.
Und DOCh kam er.
He came though.

Reiteration of old common ground information:

- (13) Hij komt toch.
Er kommt doch.
He's coming, you know he is.

Common ground marker:

- (14) Als je toch hierheen komt, neem het boek dan mee.
Falls du sowieso hierher kommst, nimm das Buch dann mit. (not with *doch*).
If you are coming here anyway, bring along the book.

Reminding causal:

- (15) Ik ben immers rijk. (not with *toch*)
Bin ich doch reich. (with obligatory inversion)
Because, as you know, I am rich.

Imperatives

Idiomatic: refusal to believe the other is sincere in what he is saying:

- (16) Kom toch!
Not in German.
Come on.

Non-idiomatic: exhortation to come, mitigating the imperative by presuming a common ground that this is the correct thing to do (?):

- (17) Kom toch!
Komm doch!
Do come!

Request for coming, while it was clear the interlocutor would not do that.

- (18) Kom TOCH!
Komm DOCH!
Change your mind and come!

Wh-questions

Reasking an already answered direct question when one has forgotten the answer:

- (19) Wie heeft er toch dat artikel over contrast geschreven?
Wer hat doch dieses Papier über Kontrast geschrieben?
Who was it that wrote this paper about contrast?
- (20) Wat was dat toch voor voetbalwedstrijd?
Was war das doch für ein Fußballspiel?
What soccer game was that?
- (21) Wie heeft er toch de cake opgegeten?
Wer hat denn den Kuchen aufgegessen? (not with *doch*)
Who ate the cake?

The meaning of *doch/toch* is not clear in the last two cases. They are only available with the unaccented *doch* and *toch*. It seems the *wh*-question must related to a common ground fact.

Exclamations

Exclamation of criticism of addressee:

- (22) Peter toch!
Aber Peter! (not with *doch*)
But Peter!

Exclamatives (surprise over CG fact?):

- (23) Wat is hij toch slim!
Wie klug er doch ist!
How clever he is!

Wishes (their fulfillment is already in the CG?):

- (24) Als hij toch zou komen!
Wenn er doch käme!
If he would come!

Conjunction

Adversative conjunction:

- (24) Die Lage ist ernst, doch nicht hoffnungslos.
Not in modern Dutch.
The situation is serious but not hopeless.

The same root in other languages

English

Proconcessive:

- (25) He is coming though.
Doch kommt er.
Toch komt hij.

Concessive conjunction:

- (26) Though he is ill, he is still coming.
Not with *toch* or *doch*.

Swedish

dokh, apparently taken from German. Mostly proconcessive.

Spoken Russian

The reminder clitic particle *-to* (McCoy 2003).

Gothic

Will be discussed in section (5).

Sanskrit

tu as the contrastive conjunction Sturtevant (1928). According to Jared Klein (p.c.) this is no longer the accepted theory about the origin of *tu*. Without Sanskrit and ignoring the Russian reminder particle, there is no evidence for a pre-Germanic origin of *doch/toch*.

3 Core meanings

As stated in the introduction, the main two assertion uses are distinguished by intonation. Without accent *toch/doch* is a modal particle for common ground status and with accent, it becomes a correction particle. These are near opposites and it seems hard to connect the two uses from a core meaning and intonation.

(27) Peter kommt DOCH.

Peter is coming and we thought he was not.

(28) Peter kommt doch.

Peter is coming as we always thought.

How to do this with intonation theories? Following Rooth (1992), the accent relates to a set of alternatives that can be obtained by replacing elements of the same category for the accented item. For Rooth, it only depends on the context which substitution instances are in the set. Zeevat (2004) argued that proper alternatives to *x* must be distinct from *x* and that distinctness requires that it is incompatible with the common ground that the two items could still be identical or overlapping. For the category of particles to which *TOCH/DOCH* belongs this would work out as conceptual incompatibility. That condition is only fulfilled by full negation, so that *Peter kommt NICHT* is the only alternative.

In the correction use, the correct alternative is given by the common ground knowledge that is to be corrected. It is of the form *p* if the utterance is DOCH NICHT *p* and of the form NICHT *p* if the utterance is DOCH *p*. But taking the meaning of the unaccented *doch* (*p* is common ground) and adding to that the salience of the negation of *p* does not give us a correction. *p* is precisely not common ground if the correction *p* makes any sense and it is not possible to assume that not-*p* is common ground if *p* is also common ground. Correction with unaccented *doch* is possible when the other speaker seems to have withdrawn his earlier commitment to *p*. *Doch* is used in this case to remind the other speaker that *p* really is common ground.

One strategy that one can follow to solve this puzzle is to assume a core meaning for *toch/doch* that underlies both the accented and the unaccented

toch/doch which derive their meaning in a given context from the core meaning and pragmatically based reasoning about the content in the context.

One of the most worked-out approaches on this line is Karagjosova (2003) that proposes a core meaning of *denial of earlier expectation*, following earlier work of Weydt (1969). In the case of correction, the earlier expectation can be equated with whatever is corrected.

(28) Peter ist also DOCh verreist.

Although I had reasons to believe that Peter would not leave, he has left.

In the non-accented case it is denial of the speaker's expectation about what the hearer believes. In the example below, B expects A to believe that Peter is away on a journey, but A's contribution can only be taken as indicating that A has forgotten all about it.

(29) A. Peter kommt also mit.

B. Er ist doch verreist.

A. Peter is coming along then.

B. But he has left, hasn't he?

In the case of the accented DOCH, one can take the salient negative alternative presupposed by the accent on DOCH as a way of indicating which expectation is violated: the expectation that Peter would not have left on a journey. And in fact one can show that there is a whole range of possible sources for the violated expectation with the common ground only being one of them: the speaker only, the hearer only, third party opinions, the linguistic context, a plausible inference from what has just been said. For the unaccented doch, the speaker's expectation that the hearer believes *p* is a plausible basis for assuming that *p* is common ground between the speaker and the hearer. If the speaker expects the hearer to believe *p* but does not in fact believe *p* herself, presumably she should indicate her dissent.

What are the problems with this account? First of all, while denial of expectation is a common ingredient of both denial of an expectation about the hearer's belief that *p* and denial of expectation that *p*, it is not clear that an intonational account on the lines of Rooth (1992) is able to relate the two expectations. Why should accent lead to denial of expectation with respect to *p*, when it is just the denial of the expectation that the hearer believes *p* in the unaccented case? This would not be a straightforward application of what is understood about accent.

The second problem is that there are subtle differences between Dutch and German: combining unaccented *doch* with questions is more frequently out, and there are cases like (29) that are not in German.

- (29) Als je toch hierheen komt, neem het boek dan mee.
If you are coming here anyway, bring along the book.
Falls du sowieso hierher kommst, nimm das Buch mit.

Now if the core meaning is the same in both languages (and there is a very large overlap) the same pragmatic reasoning should apply and there should be no such distinctions. In addition, the proposed core meaning seems to be absent in (29): it is common ground between speaker and hearer that the hearer will come to the location of the speaker and that common ground knowledge is just mobilised for planning the return of the book: there is no suggestion at all that the hearer has forgotten about his plan to come over. A similar case is the following example.

- (30) Ik ga toch 2 weken weg?
Ich bin doch weg für 2 Wochen?
As you know I am away for two weeks?

There is no indication in this example that the hearer does not know anymore about the speaker's travelling plans: the purpose of the speaker is just to bring it up again so that she can now ask the hearer to water her plants while she is away. (The absence in Dutch of the particle *ja* may explain why Dutch has a wider range of uses here: *ja* is a less ambiguous common ground marker). The reminding causal uses in German (*Bin ich doch reich.*) are however another case in point.

So while we accept in principle the possibility of a core meaning approach, we have doubts in this particular case with respect to the possibility of a core meaning theory that meets the two demands: (a) the core meaning is present in all uses and (b) each use can be fully explained using the core meaning and pragmatic reasoning only. There seems to be no proposal that fully does the job.

It is a point in favour of core meaning theories that they do justice to the intuition of a conceptual unity behind a rich variety of uses. But it does not seem that the alternative theory of a historical process in which the different uses are formed is unable to account for this intuition. A historical account has the advantage that differences between languages are not problematical and that the pragmatic reasoning becomes superfluous, or gets a different role. This is the road that we will pursue in the next section and beyond.

4 Spread in Grammaticalisation

Zeevat (2007) investigates the possibilities of simulating one central step in the grammaticalisation process in a probabilistic model: the *recruitment* of a word for a new use. In recruitment, an existing word acquires a new use, often described as weaker, more pragmatical and grammatical. Recruitment is the standard process assumed for the origin of the functional inventory of languages: all grammatical morphemes, auxiliaries, articles, prepositions, pronouns and particles, with the possible exception of demonstratives ultimately derive from lexical words by recruitment.

The model makes reproduction of a use dependent on its communicative success, i.e. on whether the hearer correctly identifies the speaker's intention and of the importance of the error (important errors lead to less reproduction). On this basis, historical events like spread and usurpation can be simulated by the change of probabilities guiding the use and interpretation of linguistic expressions for certain meanings. This makes recruitment happen only when three conditions are satisfied:

- a. the source use weakly entails the target use (if the source use obtains, the target use holds more often than not) (**push**)
- b. non-recognition of the target use leads to "serious" communicative failure
- c. non-expression of the target meaning is overwhelmingly interpreted as excluding the target use (with b together: **pull**) and the target meaning lacks an alternative expression device.

Without push, the use of the word is not able to evoke the new meaning, without pull there is no reason for the new formation.

Depending on the relative natural frequencies of the old use versus the new use, the new use can either end up coexisting with the old use (**spread**) or take over the word entirely (**usurpation**).

Usurpation is also prevented if the new use and the old use are protected from each other, i.e. there are features of both uses such that confusion of an old use for a new use or vice versa is unlikely. The attested grammaticalisation of words for “head” to become the local preposition corresponding to English *on* is a case in point, since the prepositional and noun uses cannot be confused for each other in this case.

As an example consider spread as an account for the proconcessive use of *toch/doch* as derived from its correction use:

- a. the fact that the common ground contains an opposite opinion of the speaker, the hearer or somebody else weakly entails that the common ground contains

recent information from which one may infer the opposite of what is said -- People's opinions are reasons for thinking that what they say is true

b. non-recognition leads to the availability of p in memory as a reason for thinking that not- q --- while q is the case.

(31) It rains. Peter is going for a walk.

(If it rains, people normally do not go for walks.)

Marking Peter's walking by a proconcessive is effective in removing the inference that Peter is not going for a walk.

c. A plain assertion is an answer to a fully open issue.

Compare Stalnaker's assertion conditions (Stalnaker 1979) that p is both new and consistent information, but the assumption used here is stronger: the common ground should not already contain reasons for inferring that p is true or false.

Further notice that in conversation the proconcessive use does not seem more frequent than correction and that it is hard to confuse one for the other: if there is no CG element that is corrected by the statement, the correction interpretation is out. If there is nothing in the context that normally causes the statement to be false, the proconcessive interpretation is impossible.

So spread of a correction marker to become both a correction and a proconcessive marker is possible and will happen in due time in the model unless there are other candidates for recruitment or other good ways of expressing the new meaning. It follows that it must be assumed that the other German proconcessives, e.g. *trotzdem* must have evolved as proconcessives after *doch* became one.

It is not necessary to think of spread as creating two different words just because there are two different meanings. In a study of the use of *already* as a perfective aspectual marker in Singapore English (next to its standard English use), Fong (2003) proposes the mechanism of semantic epenthesis. In this mechanism, the word projects a set of semantic features and the context can switch some of them off. Applied to this case, one could think of the correction marker spread to proconcession as projecting the combination of *proconcessive*(X) and *correction*(X), with X standing for the antecedent, so that the existence of a proconcessive antecedent switches off the correction reading or vice versa. (Notice that it is impossible to be both the negation of p and a reason for p being false, i.e. to be both the corrected item and the proconcessive

antecedent). Along these lines, a conceptual unity of the word can be maintained, even after spread.

What can spread do for *toch/doch*? Can one relate the different uses by postulating recruitment or other processes?

It is important below that the arrows are monodirectional: it should not be possible for the inverted process to happen.

It seems the uses in section (2) can be derived from each other partly by spread in the following way. This was already demonstrated for (b.) and it is plausible for (e.) as well.

- a. correction confirmation question => correction marking on anything
 - b. correction marking => proconcessive
 - c. proconcessive => contrastive conjunction
 - d. reminder question => CG marker
 - e. CG marker => mitigator (Abtönungspartikel) of imperatives, wishes, questions, exclamations, etc.
 - f. correction marker => correction particles
- (DOCH (TOCH and TOCH WEL), and DOCH NICHT (TOCH NIET))

(c.) is a different case. The proconcessive *doch* occurs clause-initially and the process underlying the formation would be a reanalysis of the antecedent and the *doch*-clause as a single sentence and of *doch* as a conjunction. (a.) and (d.) can be analysed as spread by assuming that intonation separates their use as a marker of correcting questions or reminding questions from their use as correcting or reminding assertions. The assertive force is then intonationally expressed (which epenthesises the question feature) and leaves the correcting or reminding feature. This would be a case of semantic epenthesis: the correcting confirmation question marker used on an assertion cannot mark "confirmation question" but still projects "correction". Similarly, a reminder question marker on an assertion, cannot mark "question" but "reminder" can still be projected.

This spread would result in accented *toch/doch* becoming a marker of correction and unaccented *toch/doch* in a marker of common ground.

A marker of common ground (e.) can likewise easily spread into being a mitigator. Asking somebody to do what they want to to do anyway preserves face. Once mitigation has been established, it can further spread to wishes and questions.

Finally (f.) would come out of a process of ellipsis.

It is hard to see however that there is spread that connects the accented with the unaccented *dochs*. There are just two families of uses. A curious exception is the conjunction *doch* that is unaccented and so seems to fall outside the family

of accented *dochs* to which it semantically belongs. It must be a property of conjunctions that they lose accent.

There are versions of *toch/doch* in other languages. English has *though* as a concessive conjunction and as a proconcessive, Swedish has *dokh* (adversative and concessive meaning), Russian *-to* (a reminder postclitic). These seem to be only fragments of the Dutch and German use and are not divided by accent. Gothic has a very wide range of otherwise unattested uses of *thau*, in addition to uses like the Dutch and German use.

The conclusion should be that while spread helps in accounting for the many uses (and liberates us from the task of accounting for the different uses by pragmatic reasoning from a core meaning), it still is unable to deal with the paradox with which this paper opened.

5 A Historical Explanation

In this section, we give a historical explanation of the paradox.

The particles *doch* and *toch* are derived from an Indogermanic origin *to+u+h* composed of the demonstrative *to*, the question marker *-u* and an emphatic marker *h* Hentschel (1986). Given the nature of Indogermanic question marker in which the marker *-u* can be attached to anything that is questioned, a gloss may be *That?* or *Is that so?*

It should have a role in the sentence and the most reasonable role would seem that of a question tag on an assertion: *S, is that so?*

There are three kinds of confirmation question. One can try to confirm one's own opinion, the opinion of a third party and one can confirm the opinions of the interlocutor. For the first kind, there are the so-called biased questions:

- (32) He is guilty, isn't he?
He is guilty, yes?
He is guilty, right

For confirming the opinion of a third party, an open question that quotes the opinion seems the most appropriate.

- (33) Is it true/correct that he is guilty?

The hearer's opinion is checked with rising intonation questions.

- (34) He is guilty?
He isn't guilty?
Is he guilty?

Isn't he guilty?

The Dutch and German questions in (34) have one property in addition to such rising intonation questions. They also presuppose the negation to be common ground. The surprise and the bias in (34) can also be due to the speaker's private opinion. (34) and (35) however both seek confirmation of what the hearer just said.

- (35) Hij is TOCH schuldig?
Er ist DOCH schuldig?
Is he guilty after all?

How about the unaccented versions?

- (36) Hij is toch schuldig?
Hij is toch niet schuldig?
Er ist doch schuldig?
Er ist doch nicht schuldig?

Here an old hearer/CG opinion is checked, quite possibly in reaction to some opinion of the hearer that casts doubt upon it. If so, it follows that the English question with reversed polarity is the adequate translation. But those can also be used if the speaker does not take his opinion to be common ground.

- (37) Hij is toch schuldig?
Isn't he guilty?
Hij is toch niet schuldig?
Is he guilty?

This relation with a claim of the hearer disappears when the reminder question has other reasons than seeking for confirmation as in (38)

- (38) Ik ga toch volgende week naar Spanje? Kan jij dan voor de planten zorgen?
Ich bin doch nächste Woche in Spanien? Kannst du für meine Pflanzen sorgen?
When I am in Spain next week, can you water my plants?

In both cases *to+u+h* would get tagged onto a hearer opinion and thereby the hearer would be asked to confirm it. In both cases the hearer has committed herself to *p* in the past. The speaker wants to confirm whether the commitment is valid or still valid.

In Dutch and German these uses are still there and they are the only two uses where the same pragmatic meaning (please confirm that you believe *p*) is invoked with accent (indicating an activated contrasting *not-p* and without).

In section (4) it was shown how spread can account for the other Dutch and German uses.

There has been little development in the development of *doch* and *toch* in the recent history of German or Dutch. In fact, the very similarity between the Dutch and the German particle indicates that nearly all of the uses were in place when the languages separated and that the formation of the particle is very old. If moreover the Russian *-to* really has the origin postulated, *toch* etc. predates the split between Germanic and Slavic.

The arguments that can be given for the original meaning are the etymology and especially the presence of the question marker *-u* in it. The *to* picks up the preceding sentence and the *-u* questions it. That the sentence should be a repetition of something the hearer has said or confirmed before does not follow in the same way. But it is clear that *is that so?* on its own would be just a challenge to the hearer.

Further, among the uses in German and Dutch, it appears to be the only two that can both serve as a source for all the other uses by spread and the only two that can be semantically related to each other by having or lacking contrastive stress.

It should further be noted that with the exception of Gothic none of the other languages have uses of their descendant of *to+u+h* that are not part of the Dutch/German array of uses and Gothic has a large overlap.

A weak spot is that there is no older German, older Dutch or Gothic evidence for these two original uses. This may be due to the fact that no real conversations are available for those languages and language phases. Confirmation questions typically slow down a story and the typical uncertainty about hearing it right in conversation is not part of the writing medium.

The argument requires the explanation of the Gothic uses that are not in Dutch or German.

The Gothic uses of *thau* can be resumed from the dictionary Streitberg (1910):

- a. comparative conjunction: *than* after comparative
- b. in disjunctive questions (also elliptic ones): *or*
- c. adversative conjunction: *jedoch*
- d. introducing the consequent of a conditional sentence: translation of Greek *an*
- e. pragmatic ("metacommunicative") function: in proper direct and indirect (*wh*)-questions, rhetorical (*wh*-)questions and assertions
- f. *though* as in English

Uses (c.) and (f.) are familiar from German, English and Dutch.

Comparative conjunction may be related to a different use of *to* as in the English (39).

(39) Is he that strong?

In the equivalent of *John is stronger than Bill*, the combination of *to* with *-u* would then be asking the hearer to answer the question how strong Bill is and uses that answer to make a statement about how strong John is. (Degrees to which *x* has the property *P* are often used in semantics for comparatives).

The marker of disjunctive questions can be related to examples like:

(40) skuld-u ist unsis kaisara gild giban thau ni-u?
is it right to give the emperor taxes or not?

It would seem that *thau niu* is a question tag like the Chinese *X bu X* as in (41) and that *thau* has been reanalyzed as a disjunctive question marker in this tag.

(41) Ni hao bu hao?
You good not good?
Are you doing well?

The uses in (d.) and (e.) are markers of non-veridicality and the recruitment of a question tag for this purpose seems natural. The somewhat mysterious modern uses of *toch* in dutch questions may perhaps be related to these Gothic uses. For our point, it is important that there is still a strong association with the marking of questions and non-veridicality. This makes an origin of *toch* and *doch* as question tags more plausible.

If we are right, this solves the paradox. The original use is not paradoxical, but both uses have spread a good deal. The accented *toch/doch* has become a correction and a concession marker and lost its association with questions. The unaccented *toch/doch* has become a marker for common ground information, again without an inherent relation with questions. But the new uses in the two groups cannot be related with uses in the other group as their contrastive or non-contrastive counterpart.

This merely points to the fact that accent is always assigned to a word in a particular use. Contrastive accent merely means that a contrasting element has been activated, as Rooth (1992) has it. Accent therefore plays a disambiguating role, but it does not do that as part of the inventory of sound distinctions that keep words apart in Dutch and German. It is merely a question of expressing the

presence of a contrastor, which accidentally helps to keep various uses of *toch* and *doch* apart.

Apart from contrastive accent, uses are also kept apart by the rising intonation typical of questions and by sentential position and inversion.

The solution to the paradox starts from a shared origin of the accented and the unaccented *doch*. We postulate that this is as a marker of questions with which the speaker seeks a reconfirmation of a hearer opinion. The accented *doch* indicates that, in the context, the opposite opinion is also around, often as an element of the common ground that speaker and hearer share. The absence of accent indicates that the opposite opinion is not activated. Spread created a considerable ambiguity and accent (together with syntactic factors and rising and falling intonation) helps to keep the different uses apart.

6 References

- Doherty, M. (1985). *Epistemische Bedeutung*. Akademie-Verlag, Berlin.
- Fong, V. (2003). Unmarked 'already': Aspectual expressions in two varieties of english.
- Foolen, A. (2003). Niederlaendisch *toch* und deutsch *doch*: Gleich oder doch nicht ganz? *Linguistics Online*, 13(1/3).
- Hentschel, E. (1986). *Funktion und Geschichte deutscher Partikeln*. Ja, doch, halt und eben. Tuebingen.
- Karagjosova, E. (2003). *The Meaning and Function of German Modal Particles*. Ph.D. thesis, University of Saarbruecken.
- McCoy, S. (2003). Connecting information structure and discourse structure through Kontrast: The case of colloquial Russian particles -to, ze, and ved'. *Journal of Logic, Language and Information*, 12(3), 319-335.
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1, 75-116.
- Stalnaker, R. (1979). Assertion. In P. Cole, editor, *Syntax and Semantics*, volume 9. Academic Press, London.
- Streitberg, W. (1910). *Die Gotische Bibel*. Zweiter Teil: Gotisch-griechisch-deutsches Worterbuch. Carl Winter's Universitaetsbuchhandlung, Heidelberg.
- Sturtevant, A. M. (1928). A note on the Gothic particle *thau*. *Modern Language Notes*, 43(4), 242-244.
- Weydt, H. (1969). *Abtonungspartikel*. Bad Homburg.
- Zeevat, H. (2004). Contrastors. *Journal of Semantics*, 21, 95-112.
- Zeevat, H. (2007). Simulating recruitment in evolution. In G. Bouma, I. Kramer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 175-194. KNAW, Amsterdam.

Outline of the Foundations for a Theory of Implicatures

Anton Benz

Centre for General Linguistics, Berlin

In this paper, we outline the foundations of a theory of implicatures. It divides into two parts. The first part contains the base model. It introduces signalling games, optimal answer models, and a general definition of implicatures in terms of natural information. The second part contains a refinement in which we consider noisy communication with efficient clarification requests. Throughout, we assume a fully cooperative speaker who knows the information state of the hearer. The purpose of this paper is *not* the study of examples. Our concern is the framework for doing these studies.

1 Introduction

Communication poses a coordination problem. We represent this coordination problem by signalling games (Lewis, 2002). The solutions to the coordination problem are strategy pairs which describe the speaker's signalling and the hearer's interpretation behaviour. The behaviour is an objective natural regularity, and the speaker's and hearer's strategies determine with which probability they will choose their respective actions given their respective information states. As natural regularity, the communicative process can be described as a causal Bayesian network (Pearle, 2000). From this representation, we derive the notion of *natural information* which is related to Grice' (1957) concept of *natural meaning*. We claim that this is a key concept for the understanding of pragmatics.

Natural information is *objective* information, i.e. it exists independently of the beliefs and intentions of language users. To justify this interpretation we have to interpret the probabilities in signalling games as *objective relative frequencies*. From this objective level we distinguish a subjective cognitive level at which probabilities are interpreted as *subjective probabilities*. We describe the

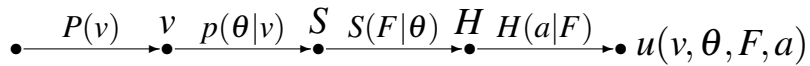
subjective level by *optimal answer* (OA) models. We justify this representation by a discussion of the theory of mind as incorporated in *iterated best response models* (Franke, 2009).

Accordingly, the first part of the paper divides into five sections. The first section introduces signalling games, the second section the concept of natural information and the general definition of implicature, and the third section the optimal answer models and their *canonical* solutions. The third section also discusses the relation between OA and iterated best response models. The fourth section applies the general definition of implicatures to OA models and signalling games. In Section 2, we present a lemma which provides us with a criterion for deciding whether or not a strategy pair is an objective Pareto Nash equilibrium of a signalling game. This lemma, Lemma 2.3 will play an important role in our discussion of aspects of bounded rationality, the theory of mind, and the *objective* justification of canonical solutions to OA models. The last section of the first part provides the proof of this lemma.

The second part of this paper starts out with a discussion of the idea that ambiguities are resolved by choosing the more probable interpretation, and that, as a consequence, the more probable interpretation of an ambiguous utterance is communicated with *certainty*. This principle figures prominently in Prashant Parikh's (2001) approach to game theoretic pragmatics, which basically assumes that all pragmatic strengthening and weakening of interpretation can be reduced to cases of disambiguation. We argue that the natural hearer's reaction to an ambiguity is to ask a clarification request. Hence in Section 8, we consider signalling games for which the hearer's action set contains efficient clarification requests. *Efficiency* means that clarification requests have nominal costs and lead to almost maximal payoffs. The availability of efficient clarification requests changes the equilibria of signalling games if we allow for *noisy* speaker strategies. This noise may have *external* causes, i.e. the kind of noise might not be predictable from game theoretic parameters. Hence, we introduce a very general model for representing noisy speaker strategies. This is done in Section 9. In this section, we also show how the canonical solutions to OA models change, and how the notion of implicatures applies to models representing noisy speaker strategies. Section 10, contains further characterisations of the equilibrium properties of canonical solutions for noisy games and the proof of a lemma analogous to Lemma 2.3. The final section contains some clarifications concerning our concept of *nominal* costs.

2 Signalling Games

Grice (1989, p. 26) characterised conversation as a *cooperative effort*. This means that the contributions of the interlocutors are not isolated sentences but subordinated to a joint purpose. In this paper, we will always assume that each assertion answers an implicit or explicit question by the hearer which in turn is embedded in a decision problem. The decision problem is such that the hearer has to make a choice between several actions. The hearer's choice of actions depends on his preferences regarding the actions' outcomes and his knowledge about the world. The speaker's message helps the inquirer in making his choice. The quality of a message depends on the action to which it will lead. Hence, communication poses a coordination problem to speaker and hearer. The speaker has to choose his contribution such that it induces the hearer to choose an optimal action; and the hearer has to consider the speaker's message and use the communicated information for making the best choice. We represent these coordination problems as *signalling games* (Lewis, 2002). The signalling games are such that first nature chooses a world v with probability $P(v)$; then again nature chooses a type θ , i.e. an information state, for the speaker S with conditional probability $p(\theta|v)$; then the speaker chooses a signal F with conditional probability $S(F|\theta)$, and finally the hearer chooses an act a with conditional probability $H(a|F)$. A branch of this game is depicted in the following figure:



We formally define the signalling games as follows:

Definition 2.1 (Signalling Game) A tuple $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ is a signalling game if:

1. Ω and Θ are non-empty finite sets;
2. $P(\cdot)$ is a probability distribution over Ω ;
3. $p(\cdot|v)$ is a probability distribution over Θ for every $v \in \Omega$;
4. \mathcal{F} and \mathcal{A} are respectively the speaker's and hearer's action sets;
5. $u : \Omega \times \Theta \times \mathcal{F} \times \mathcal{A} \rightarrow \mathbb{R}$ is a shared utility function.

We assume that $u(v, \theta, F, a)$ can be decomposed into a difference $u(v, a) - c(F)$ for some real valued function $u(v, a)$ and a positive value $c(F)$.

We assume that the general game structure is common knowledge. The speaker, in addition, knows θ when choosing signal F , and the hearer knows F when choosing action a . This means that the agents' strategies are functions of the following form:

- For each type $\theta \in \Theta$, the speaker's strategy $S(\cdot|\theta)$ is a probability distribution over \mathcal{F} ;
- For each signal $F \in \mathcal{F}$, the hearer's strategy $H(\cdot|F)$ is a probability distribution over \mathcal{A} .

In principle, the probabilities could be interpreted as objective frequencies or as subjective probabilities. For reasons which will become clear in the next section, we interpret all the probabilities related to signalling games as objective frequencies.

Next, we introduce the notion of a *Nash equilibrium*. The speaker's expected utility $\mathcal{E}(S|H)$ of strategy S given a hearer strategy H is defined as:

$$\mathcal{E}(S|H) = \sum_{v \in \Omega} P(v) \sum_{\theta \in \Theta} p(\theta|v) \sum_{A \in \mathcal{F}} S(F|\theta) \sum_{a \in \mathcal{A}} H(a|F) u(v, \theta, F, a). \quad (2.1)$$

As the basic signalling games defined in Def. 2.1 are games of pure coordination, i.e. games in which the utility functions of both agents are identical, it follows that $\mathcal{E}(S|H) = \mathcal{E}(H|S)$. With these notions at hand, we can define:

Definition 2.2 (Nash Equilibrium) A strategy pair (S, H) is a Nash equilibrium of a signalling game $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ iff:

1. For all speaker strategies S' : $\mathcal{E}(S'|H) \leq \mathcal{E}(S|H)$,
2. For all hearer strategies H' : $\mathcal{E}(H'|S) \leq \mathcal{E}(H|S)$,

The equilibrium is strict if we can replace \leq by $<$. It is weak if it is not strict.

For a game of pure coordination, a Nash equilibrium is a *Pareto Nash equilibrium* iff for all other Nash equilibria (S', H') : $\mathcal{E}(S'|H') \leq \mathcal{E}(S|H)$. In this case, we also say that (S, H) (weakly) Pareto dominates (S', H') .

The textbook equilibrium concept for signalling games is the concept of a *Bayesian perfect equilibrium*. Bayesian perfection takes the player's information set into account. The player's strategy must be optimal given the information available to him at the time when he actually makes the decision. For the hearer, this is after receiving an answer F . Apart from the possible semantic meaning of the answer, the hearer is gaining additional information from the fact that the answer was given. Hence, the probability distribution that enters in the hearer's decision making is his prior distribution updated with the information gained by learning that a certain answer has been given. But, for the basic

signalling games which we consider, Bayesian perfect equilibria and Nash equilibria in the sense of Definition 2.2 coincide. Although their definition is more complicated, it can be easier to do calculations for Bayesian perfect equilibria. We will do this in Section 6.

In general, it is often convenient or necessary to formulate constraints and do calculations with conditional probabilities, and not with P and p directly. The probability with which nature assigns type θ to speaker S in world v equals $P(v) p(\theta|v)$. Hence, the speaker's probability $\mu_s(v|\theta)$ for a world v after receiving type θ is a conditional probability defined as the probability to receive θ in v divided by the overall probability of receiving θ ; see (2.2). For the hearer, we find an analogous probability distribution. He acts after receiving a signal F . Hence, the hearer's probability $\mu_h(v|F)$ of a world v after receiving F is the probability of receiving F in v divided by the overall probability of receiving signal F (2.2). The explicit definitions are as follows:

$$\mu_s(v|\theta) = \frac{P(v) p(\theta|v)}{\sum_w P(w) p(\theta|w)}, \quad \mu_h(v|F) = \frac{P(v) \sum_{\theta} p(\theta|v) S(F|\theta)}{\sum_w P(w) \sum_{\theta} p(\theta|w) S(F|\theta)}. \quad (2.2)$$

Here and in the following, we assume that the denominators are non-zero. For μ_s this means that there exists a w such that $P(w) p(\theta|w) > 0$, and for μ_h that there are w and θ for which $P(w) p(\theta|w) S(F|\theta) > 0$.

In later sections, we will often make use of the following abbreviations:

$$\mu_{\Theta}(\theta) := \sum_w P(w) p(\theta|w), \text{ and } \mu_{\mathcal{F}}(F) := \sum_w P(w) \sum_{\theta} p(\theta|w) S(F|\theta). \quad (2.3)$$

$\mu_{\mathcal{F}}(F)$ is the probability for the speaker producing F , and $\mu_{\Theta}(\theta)$ is the probability for the speaker's type to be θ . As it is clear from the argument which measure is meant, we will write $\mu(F)$ instead of $\mu_{\mathcal{F}}(F)$, and $\mu(\theta)$ instead of $\mu_{\Theta}(\theta)$.

Given type θ , the (speaker's) *expected utility* of an action a is defined by:

$$\mathcal{E}_s(a|\theta) = \sum_v \mu_s(v|\theta) u(v, a) \quad (2.4)$$

Similarly, given answer F , the (hearer's) *expected utility* of an action a is defined by:

$$\mathcal{E}_h(a|F) := \sum_v \mu_h(v|F) u(v, \theta, F, a). \quad (2.5)$$

The speaker's expected utility of a strategy S given his type θ is then:

$$\mathcal{E}_s(S|\theta) = \sum_A S(F|\theta) \sum_a H(a|F) \mathcal{E}_s(a|\theta) \quad (2.6)$$

And the hearer's expected utility of a strategy H given his information state after receiving signal F is then:

$$\mathcal{E}_H(H|F) := \sum_a H(a|F) \mathcal{E}_H(a|F) \quad (2.7)$$

We are now interested in a simple criterion for deciding whether a strategy pair is a Pareto Nash equilibrium. The criterion will only depend on S , H and the following set $\mathcal{B}(\theta)$ which is the set of all actions with maximal expected utility:

$$\mathcal{B}(\theta) = \{a \in \mathcal{A} \mid \forall b \in \mathcal{A} \mathcal{E}_S(b|\theta) \leq \mathcal{E}_S(a|\theta)\}. \quad (2.8)$$

Throughout the paper, we will make extensive use of the following fundamental lemma:

Lemma 2.3 *Let $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a signalling game. Let Θ^* be the set of all types θ for which $\exists v P(v) p(\theta|v) > 0$. Let (S, H) be a strategy pair which satisfies the following condition:*

$$\forall F \in \mathcal{F} \forall \theta \in \Theta^* (S(F|\theta) > 0 \Rightarrow H(\mathcal{B}(\theta)|F) = 1). \quad (2.9)$$

Then (S, H) is a Pareto Nash equilibrium. Furthermore, if H' is such that

$$\exists F \in \mathcal{F} \exists \theta \in \Theta^* \exists a \notin \mathcal{B}(\theta) (S(F|\theta) > 0 \wedge H'(a|F) > 0), \quad (2.10)$$

Then (S, H') is not a Nash equilibrium, in particular, it is $\mathcal{E}(H'|S) < \mathcal{E}(H|S)$.

We will prove this lemma in Section 6

3 Natural Information

In (1957), Grice introduced the distinction between *natural meaning* and *communicated meaning*. Natural meaning is the information which can be carried by an event or object independently of the beliefs and intentions of any person who may use this event or object for the purposes of communication. Grice used the following example for illustrating the concept of *natural meaning*:

- (1) a) Those spots mean measles.
- b) Those spots didn't mean anything to me, but to the doctor they meant measles.

In both sentences, the word *meaning* refers to natural meaning. The spots carry the information that the patient is infected with measles independently of any person using the spots for communicating that he is infected with measles, e.g. by pointing at the patient and saying: '*Look what he has!*' The spots carry their

information due to a causal relation that exists between the infection and red spots on the skin. This causal relation is a natural regularity which is the basis for the inference from *red spots* to *measles*.

Causal relations can be represented by *causal networks*. The diagram in Figure 1 from (Pearle, 2000, p. 15) may serve as an illustration. $\mathcal{X}_0, \dots, \mathcal{X}_4$

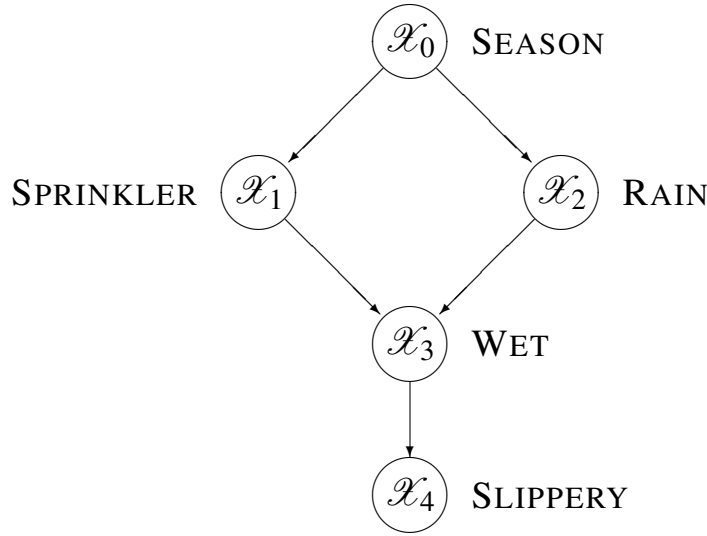


Fig. 1: A causal network.

are random variables which represent the state of the season and of a sprinkler, whether it rains, and whether a certain place is wet or slippery. The random variable for the season can take four different values, whereas the random variables for the sprinkler, the rain, and the wetness and slipperiness are only taking the Boolean values *true*, or *false*. In causal Bayesian networks, the causal dependencies are represented by *conditional probabilities* which hold between random variables. Given, e.g., that the slipperiness of a road is determined by its wetness, which in turn is determined by the fact whether a sprinkler is on, or whether it is raining, and that for example the state of the sprinkler is determined by the season, then we could say that: ‘*That the street is slippery means that the sprinkler was on or that it rained;*’ or ‘*That the sprinkler is on means that it is summer*’. In both cases, the word *means* refers to natural meaning.

We now turn to the communication process. As we have seen in the last section, the context of communication can be described by the state of the world v , the speaker’s information state θ , and a fixed information state of the hearer. Let Ω be the set of all possible worlds, and Θ of all possible speaker states. Again as in the last section, we identify the communicative behaviour of speaker and hearer with strategies S and H , i.e. with functions S which map the speaker’s possible information states θ to probability distributions over a set \mathcal{F} of possible utterances, and functions H which map utterance F to probability distributions over a set of hearer actions \mathcal{A} . Hence, S only depends on the speaker’s

information state θ , and the hearer's strategy on the signal F which he receives from the speaker. We write $P(v)$ for the probability of a world v , and $p(\theta|v)$ for the probability of the speaker's information state θ given v . If P , p , S , and H are given, then we can think of the communicative process as a *Markovian* process, i.e. a process in which the probability of each successor state only depends on the predecessor states. A branch in this process is shown in the following graph:

$$\begin{array}{ccccccc} v & & S & & H & & a \\ \bullet & \xrightarrow{P(v)} & \bullet & \xrightarrow{p(\theta|v)} & \bullet & \xrightarrow{S(A|\theta)} & \bullet \\ & & \theta & & A & & \end{array}$$

In generally, we can think of the Ω , Θ , \mathcal{F} , and \mathcal{A} as random variables in a causal Bayesian network in which the conditional probabilities P , p , S , and H define causal dependencies between these variables. Clearly, this identification assumes that all probabilities are objective frequencies. This is all we need to introduce a meaningful definition of *natural information*.

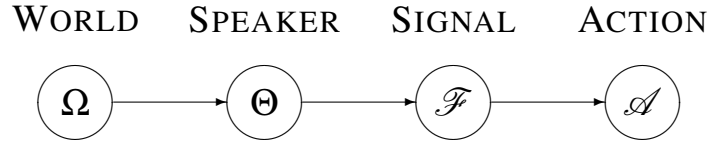


Fig. 2: The causal network associated to a signalling game.

For the following definitions, we abstract away from all particularities of linguistic communication. In order to make our definition not too far removed from our applications, we consider only graphs which represent a linear sequence of causal dependencies. But our definitions will immediately generalise to any causal Bayesian network which is represented by a directed acyclic graph. A linear graph of length $n+1$ is given by a pair $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ for which:

1. $(\mathcal{X}_i)_{i=0, \dots, n}$ is a family of non-empty sets,
2. $p_0(\cdot)$ a probability distribution over \mathcal{X}_0 ,
3. for $i > 0$ and $x_{i-1} \in \mathcal{X}_{i-1}$, $p_i(\cdot|x_{i-1})$ is a conditional probability distribution over \mathcal{X}_i .

We call a pair $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ a *linear causal network*.

From the p_i 's we can define the *joint* distributions μ^k on the product space $\mathcal{X}^k := \prod_{i=0}^k \mathcal{X}_i$, $k \leq n$, by

$$\mu^k(x_0, \dots, x_k) := \prod_{i=0}^k p_i(x_i|x_{i-1}). \quad (3.11)$$

We write μ for μ^n . As for each sequence $\mathbf{x} = \langle x_0, \dots, x_n \rangle \in \mathcal{X}^n$ the probability of x_{i+1} does only depend on its predecessor x_i , the processes defined

by $(\mathcal{X}_i, p_i)_{i=1, \dots, n}$ has the general properties of a *Markovian* processes (Pearle, 2000, p. 14).

We are now going to introduce the *marginal* probabilities. Let π_i denote the *projection* of \mathcal{X}^k onto \mathcal{X}_i ; i.e. for $i \leq k$ and $\mathbf{x} = \langle x_0, \dots, x_k \rangle \in \mathcal{X}^k$ let $\pi_i(\mathbf{x}) := x_i$, and for $X \subseteq \mathcal{X}^k$ let $\pi_i(X) = \{\pi_i(\mathbf{x}) \mid \mathbf{x} \in X\}$. For $X \subseteq \mathcal{X}_i$ we set

$$\pi_i^{-1}[X] := \{\mathbf{x} \in \mathcal{X}^n \mid \pi_i(\mathbf{x}) \in X\}. \quad (3.12)$$

We define the *marginal* probabilities μ_i on \mathcal{X}_i by:

$$\mu_i(X) = \mu(\pi_i^{-1}[X]), \text{ for } X \subseteq \mathcal{X}_i. \quad (3.13)$$

For $i \leq k \leq n$, $X \subseteq \mathcal{X}_i$, it holds $\mu^k(\pi_i^{-1}[X]) = \mu^n(\pi_i^{-1}[X])$. Hence, the definition of the marginal probabilities μ_i in (3.13) does not depend on the fact that it is defined relative to μ^n . By induction it can be shown that $\mu_i(X)$ equals

$$\sum_{x_0 \in \mathcal{X}_0} p_0(x_0) \sum_{x_1 \in \mathcal{X}_1} p_1(x_1|x_0) \dots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} p_{i-1}(x_{i-1}|x_{i-2}) \sum_{x_i \in X} p_i(x_i|x_{i-1}) \quad (3.14)$$

Finally, we define *conditional marginal probabilities* $\mu_{i|j}$ as follows: let $X \subseteq \mathcal{X}_i$, and $Y \subseteq \mathcal{X}_j$ with $\mu_j(Y) > 0$, then the conditional marginal probability of X given Y is defined by:

$$\mu_{i|j}(X|Y) = \mu(\pi_i^{-1}[X] \mid \pi_j^{-1}[Y]). \quad (3.15)$$

With these preparations, we can introduce our general definition of *natural meaning*:

Definition 3.1 Let $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ be a linear causal network. Then, for $X \subseteq \mathcal{X}_i$ and $Y \subseteq \mathcal{X}_j$ with $\mu_j(Y) > 0$, we set

$$(\mathcal{X}_i, p_i) \models Y \Rightarrow X : \iff \mu_{i|j}(X|Y) = 1. \quad (3.16)$$

We say that event Y naturally means that X .

If all \mathcal{X}_i are countable, then there is a smallest set X which is naturally implied by the occurrence of an event Y . We can identify this set with the *the natural meaning* of Y .

If X and Y are singletons, i.e. if $X = \{x\}$ and $Y = \{y\}$, then we write $\mu_{i|j}(x|y)$ instead of $\mu_{i|j}(\{x\}|\{y\})$. Furthermore, if i and j are clear from context, e.g. because x can only be an element of \mathcal{X}_i , or X a subset of \mathcal{X}_i , then we write μ instead of μ_i , or $\mu_{i|j}$.

In (3.16), nothing depends on the fact that $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ is a linear causal network. The p_i s could equally well depend on any set of random variables \mathcal{X}_j as long as $j < i$. But the condition of linearity plays an important role if we apply the concept of *natural meaning* to signalling games. Here, the fact that

signalling games in the sense of Definition 2.1 define linear causal networks entails that the *common natural information* of speaker and hearer is identical to the hearer's information state! We show this in Lemma 3.4 at the end of this section.

We introduce the relevant notion of *common natural information* in full generality. Let $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ be given. For $\mathbf{x} \in \mathcal{X}^n$ and $I \subseteq \{0, \dots, n\}$ let $\mathbf{x}|_I$ be the restriction of \mathbf{x} to I , i.e. it is the function with domain I and values $(\mathbf{x}|_I)(i) = \pi_i(\mathbf{x})$. We set:

$$[\mathbf{x}|_I] := \{\mathbf{y} \in \mathcal{X}^n \mid \mu(\mathbf{y}) > 0 \wedge \mathbf{x}|_I = \mathbf{y}|_I\}. \quad (3.17)$$

For $\mathbf{x} \in \mathcal{X}^n$ we define the common natural information by the following construction:

$$\begin{aligned} E_{I,J}(\mathbf{x}) &= [\mathbf{x}|_I] \cup [\mathbf{x}|_J], \\ E_{I,J}^0(\mathbf{x}) &= \{\mathbf{x}\}, \\ E_{I,J}^{n+1}(\mathbf{x}) &= \bigcup \{[\mathbf{y}|_I] \cup [\mathbf{y}|_J] \mid \mathbf{y} \in E_{I,J}^n(\mathbf{x})\}, \\ \text{CNI}_{I,J}(\mathbf{x}) &= \bigcup_n E_{I,J}^n(\mathbf{x}). \end{aligned} \quad (3.18)$$

The index sets I and J represent the information states of two agents. Hence, $\text{CNI}_{I,J}(\mathbf{x})$ corresponds to the standard definitions of *common knowledge*. *Implied* information is generally considered to be part of the common knowledge. As we explicate implicatures as common natural information, we have to spell out what it means that an event Y carries the information that an event X is common natural information. Hence, let $Y \subseteq \mathcal{X}_j$, $X \subseteq \mathcal{X}_i$, and $\mathbf{x} \in \mathcal{X}^n$. We obviously have to conditionalise the conditional marginal probability in (3.16) to $\text{CNI}_{I,J}(\mathbf{x})$; i.e. we have to replace the condition $\mu(\pi_i^{-1}[X]|\pi_j^{-1}[Y]) = 1$ by the condition $\mu(\pi_i^{-1}[X]|\pi_j^{-1}[Y] \cap \text{CNI}_{I,J}(\mathbf{x})) = 1$. First, if this definition should capture the common natural information carried by event Y for two agents represented by the index sets I and J , then Y should be known to both of them, hence, it should hold that $j \in I \cap J$. Second, from this it follows that the condition is reasonable only if $\pi_j(\mathbf{x}) \in Y$. These two restrictions entail that $\mu(\pi_i^{-1}[X]|\pi_j^{-1}[Y] \cap \text{CNI}_{I,J}(\mathbf{x})) = \mu(\pi_i^{-1}[X]|\text{CNI}_{I,J}(\mathbf{x}))$. Hence, the definition of common natural information for a branch \mathbf{x} cannot depend on the set Y of observable values. This straightforwardly leads to the following definition of an event X being common natural information for a branch \mathbf{x} and agents represented by index sets I, J :

Definition 3.2 Let $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ be a linear causal network. Then, for $X \subseteq \mathcal{X}_i$, $\mathbf{x} \in \mathcal{X}^n$ with $\mu(\mathbf{x}) > 0$, we set for $I, J \subseteq \{0, \dots, n\}$, $I, J \neq \emptyset$:

$$(\mathcal{X}_i, p_i, \mathbf{x}) \models \text{C}_{I,J}X : \iff \mu(\pi_i^{-1}[X]|\text{CNI}_{I,J}(\mathbf{x})) = 1. \quad (3.19)$$

We apply these notions to signalling games as follows: For a given signalling game, we identify \mathcal{X}_0 with Ω , \mathcal{X}_1 with Θ , \mathcal{X}_2 with \mathcal{F} , and \mathcal{X}_3 with \mathcal{A} ; accordingly, $p_0 = P$, $p_1 = p$, $p_2 = S$, and $p_3 = H$. The information states of the interlocutors are $I = \{1, 2\}$ for the speaker and $J = \{2\}$ for the hearer. A branch in the product space \mathcal{X}^3 is a sequence $\mathbf{b} = \langle v, \theta, F, a \rangle$. We simplify notation and write $\mathbf{b}(\Omega)$, $\mathbf{b}(\Theta)$, $\mathbf{b}(\mathcal{F})$, and $\mathbf{b}(\mathcal{A})$ instead of $\pi_0(\mathbf{b})$, $\pi_1(\mathbf{b})$, etc.

In signalling games it holds that the hearer's information state J is a subset of the speaker's information state I . This leads to a significant simplification of (3.19). First, we note that it obviously holds that:

$$J \subseteq I \Rightarrow [\mathbf{x}|_I] \subseteq [\mathbf{x}|_J]. \quad (3.20)$$

Furthermore, by induction it can be shown that:

$$i \in I \cap J \Rightarrow \forall n > 0 \forall \mathbf{y} \in E_{I,J}^n(\mathbf{x}) \pi_i(\mathbf{y}) = \pi_i(\mathbf{x}). \quad (3.21)$$

From these two facts, it follows by induction that $J \subseteq I$ implies that $\forall n > 0 E_{I,J}^n(\mathbf{x}) = [\mathbf{x}|_J]$, and hence that:

$$J \subseteq I \Rightarrow \text{CNI}_{I,J}(\mathbf{x}) = [\mathbf{x}|_J]. \quad (3.22)$$

Identifying *implicatures* of an utterance F with the common natural information carried by this event, we arrive at:

Definition 3.3 (Implicature) *Let (S, H) be a strategy pair for a signalling game $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$. Let $(\mathcal{X}_i, p_i)_{i=0, \dots, n}$ be the linear causal network defined by identifying \mathcal{X}_0 with Ω , \mathcal{X}_1 with Θ , \mathcal{X}_2 with \mathcal{F} , and \mathcal{X}_3 with \mathcal{A} ; accordingly, $p_0 = P$, $p_1 = p$, $p_2 = S$, and $p_3 = H$. Let $X \subseteq \mathcal{X}_i$, $I = \{1, 2\}$ and $J = \{2\}$. Let μ be the probability distribution on the product space \mathcal{X}^3 defined in (3.11), and let \mathbf{b} be a branch in \mathcal{X}^3 with $\mu(\mathbf{b}) > 0$. Then we set for $\mathbf{b}(\mathcal{F}) = F$:*

$$\langle \mathcal{G}, S, H, \mathbf{b} \rangle \models F +> X : \iff (\mathcal{X}_i, p_i, \mathbf{b}) \models \text{C}_{I,J} X. \quad (3.23)$$

We then say that in \mathbf{b} the utterance of F implicates that X . We simply say that the utterance of F implicates that X , $\langle \mathcal{G}, S, H \rangle \models Y +> X$, if $\langle \mathcal{G}, S, H, \mathbf{b} \rangle \models F +> X$ for all \mathbf{b} for which $\mathbf{b}(\mathcal{F}) = F$ and $\mu(\mathbf{b}) > 0$. Then, for $Y \subseteq \mathcal{F}$, we generalise:

$$\langle \mathcal{G}, S, H \rangle \models Y +> X : \iff \forall F \in Y \langle \mathcal{G}, S, H \rangle \models F +> X. \quad (3.24)$$

According to the generalisation in (3.24), a set Y of signals implicates X if every form $F \in Y$ implicates X . By (3.22), it immediately follows that:

Lemma 3.4 *Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a signalling game, and (S, H) a strategy pair for \mathcal{G} . Let $\mu_{i|\mathcal{F}} := \mu_{i|2}$ be the conditional marginal probability distribution defined in (3.15) for the linear causal network $(\mathcal{X}_i, p_i)_{i=0, \dots, 3}$ defined*

by $\langle \mathcal{G}, S, H \rangle$. Then, for $X \subseteq \mathcal{X}_i$, $Y \subseteq \mathcal{F}$, it holds:

$$\langle \mathcal{G}, S, H \rangle \models Y +> X \iff \mu_{i|\mathcal{F}}(X|Y) = 1 \quad (3.25)$$

In the following, we will often identify a solved signalling game $\langle \mathcal{G}, S, H \rangle$ with its associated linear causal network $(\mathcal{X}_i, p_i)_{i=0,\dots,3}$ and write e.g. $\langle \mathcal{G}, S, H \rangle \models Y \Rightarrow X$ iff $(\mathcal{X}_i, p_i)_{i=0,\dots,3} \models Y \Rightarrow X$ in the sense of Def. 3.1. Using this convention, we can rewrite (3.25) equivalently as

$$\langle \mathcal{G}, S, H \rangle \models Y +> X \iff \langle \mathcal{G}, S, H \rangle \models Y \Rightarrow X, \quad (3.26)$$

i.e. Y *implies* X iff Y *naturally means* X .

We further explore the potential of Definition 3.3 in Section 5.

4 The Solution Concept

4.1 Preliminary Remarks

With the terminology of Section 3, the conditions of Lemma 2.3 can now be reformulated as follows: If $\langle \mathcal{G}, S, H \rangle$ is such that an utterance of F *naturally means* that the hearer chooses a speaker optimal act, then (S, H) is a Pareto Nash equilibrium; if $\langle \mathcal{G}, S, H \rangle$ is such that an utterance of F does *not* naturally mean that the hearer chooses a speaker optimal act, then (S, H) is *not* a Pareto Nash equilibrium. We mentioned before that we interpret the probabilities in signalling games as objective probabilities. Hence, Lemma 2.3 provides us with a criterion for deciding whether a strategy pair is an *objective* Pareto Nash equilibrium.

In principle, there are two interpretations of probabilities which are of interest to us: the interpretation as objective frequencies, and the interpretation as subjective probabilities in the sense of (Savage, 1972). We will use both interpretations depending on which aspect of communication we are modelling. We interpret probabilities objectively if we want to explain the objective success of communication seen as a real world phenomenon; we interpret them subjectively if we model the cognitive level. Objective probabilities are just the familiar relative frequencies. Subjective probabilities are mathematical constructs which offer concise representations of the agent's propensities for choosing actions; i.e. assigning subjective probability P_X and utility function u_X to agent X means that X 's preferences over actions a after learning F are indistinguishable from an agent's preferences who chooses between actions according to the expected utilities $EU_X(a|F)$. As subjective probabilities are mathematical constructs, assigning them to agents does not mean that these agents actually represent these probabilities, or reason with them. Likewise, subjective probabilities do, in general, not have to correspond to observable frequencies. Objective frequencies may be completely unknown to our interlocutors; it may even

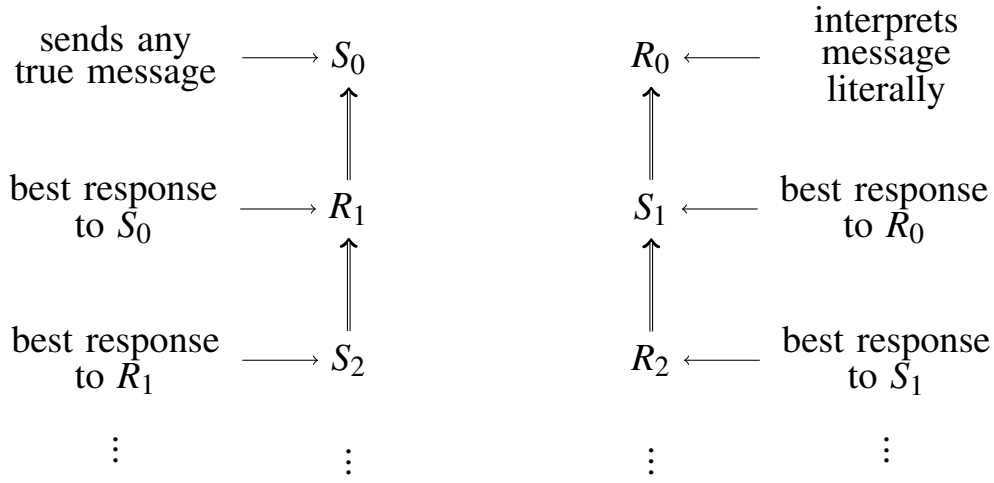
be that they don't even possess a notion of *probability*. As the probabilities P and p defined in signalling games represent the probabilities with which *nature* is choosing worlds and speaker's types, they have to be interpreted as objective frequencies, hence they might not be known to the interlocutors. In this section, we provide a model of the communicative situation which only represents the interlocutors' subjective expectations about the state of the world but not the objective frequencies with which nature chooses the world or the speaker's type.

The task is to describe the communicative situation in terms of its cognitively relevant parameters, and to provide a method for finding solutions (S, H) to the coordination problem posed by the communicative situation. As our models are intended as models of online communication, it is *prima facie* reasonable to look for a method which is *as simple as possible*.

In most game theoretic models, equilibrium concepts are describing the stable patterns of behaviour which can emerge from the interaction of rational agents in certain classes of games. As different populations playing these games may adopt different behaviours, the task in empirical applications is to find the set of all possible strategy profiles which satisfy a given equilibrium concept and to show that the behavioural patterns found in the different populations correspond to one or the other strategy profile in this set. In this paper, we follow a different strategy. We assume that there is a signalling strategy established in the population which defines the *semantic* meaning of signals (Lewis, 2002); i.e. we assume that the speaker's signals have a predefined meaning which restricts their use. The pure semantic meaning of signals also defines a hearer strategy for choosing between available actions after learning the signal's semantic meaning. Starting out from this situation, we are interested in the Nash equilibrium (S, H) which is *closest* to the given semantic convention. We think of the *distance* in terms of the number of steps of reasoning about each other which are involved in reaching the equilibrium. This can be made more precise in the framework of *iterated best response* (IBR) models (Jäger and Ebert, 2009; Franke, 2009).¹ IBR models explicate the reasoning about each other by an iterated process. In each step of this process, one of the two interlocutors chooses a best response strategy to the strategy which he assumes the other interlocutor has chosen in the previous step. There are two possible strategies from which the IBR process can start: the process can either start with a speaker strategy or with a hearer strategy. Accordingly, the model consists of two separate lines of reasoning. These two lines are shown in Figure 3.

In the IBR models worked out by (Jäger and Ebert, 2009; Franke, 2009), the S_i and R_i are in fact *sets* of strategies. In (Franke, 2009), S_0 is the set of

¹The following sketch of the IBR model is a simplified version of (Franke, 2009). For more details, motivation, and differences between the models, we refer to the original papers.



in one line have no influence on the strategy sets in the other line. Hence, let us consider the line starting with the speaker strategies in S_0 . The hearers set of best responses R_1 will in general be different from R_0 as the fact that a signal was sent may carry information in addition to the semantic meaning of the signal. As the strategies in S_0 randomly produced true signals, S_2 , the speaker's best responses to R_1 , will in general be different to S_0 . Hence, a stable state cannot be reached before S_2 is reached. The earliest stage at which the hearer can see that he has reached a stable state is therefore the stage in which he calculates R_3 ; and the earliest stage at which the speaker can see that he has reached a stable state is, accordingly, the stage in which he calculates S_4 . Hence, for the line starting with S_0 , for reaching a stable state, the hearer must at least consider the speaker's best response to his best response to the speaker's random strategy; and the speaker has at least to consider the hearer's best responses to the speaker's best responses to the hearer's best responses to the speaker's random strategies in S_0 . Let us now turn to the line of the IBR model starting with R_0 . The earliest stage at which the hearer can see that he has reached a stable state is the stage in which he calculates R_2 ; and the earliest stage at which the speaker can see that he has reached a stable state is, accordingly, the stage in which he calculates S_3 . Hence, for the line starting with R_0 , the hearer must at least consider the speaker's best response to his basic strategies in R_0 , and the speaker has at least to consider the hearer's best responses to the speaker's best responses to the hearer's basic strategies. As R_0 is, in general, not identical to R_1 , the speaker's set S_1 of best responses to R_0 will, in general, also be different from S_2 . Hence, if one line stops at an early stage, it is no guarantee that the other line does also stop early. If we take the IBR model serious as a cognitive model, then these reasoning steps must be a cognitive reality. In this section, we show that the coordination problem posed by communication can be solved with fewer steps of reasoning about each other than predicted by the IBR model. More precisely, we show that backward induction provides a solution which guarantees that speaker and hearer have reached a stable strategy pair without having to calculate *whether they have reached a stable state*.

The IBR model shows that, in order to find out whether a strategy is stable by reasoning about each other, the hearer must take into account the speaker's best response to a hearer strategy at least once. Hence, the shortest possible path to a stable strategy is the $R_0-S_1-R_2-S_3$ -path. If the method for finding a stable solution should be *simpler* or *shorter* than the method provided by the IBR model, then we have to find a method which avoids some steps of reasoning about each other in this sequence. In this respect, the simplest method is backward induction. When applying backward induction to a signalling game \mathcal{G} , the hearer does never consider the speaker's strategy, and the speaker considers the hearer's strategy only once. This is the cognitively least demanding method

for finding solutions. We will show in Section 4.3 that the resulting strategy pair (S, H) guarantees that for any possible utterance the signal *naturally means* that the hearer chooses a speaker optimal act. From Lemma 2.3 it follows that (S, H) is a Pareto Nash equilibrium; hence it is a stable strategy pair. There is no need for further steps of reasoning about each other. The following method for finding a solution to the coordination problem described by signalling games was introduced in (Benz, 2006). We call it *the Optimal–Answer (OA) model*.

4.2 The Optimal–Answer Model

In this section, the general features of the communicative situation are the same as that considered in the context of signalling games. We again assume that the conversation is subordinated to a joint purpose which is defined by a decision problem of the hearer. This decision problem may be revealed by an implicit or explicit question by the hearer. Hence, we can call the speaker’s message an *answer*. The OA model tells us which answer a rational language user will choose given the hearer’s decision problem and his knowledge about the world. We call the basic models which represent the utterance situation as *support problems*. They consist of the hearer’s decision problem and the speaker’s expectations about the world. These expectations are represented by subjective probabilities. In (Benz, 2006, 2007), it was shown that, in general, it is not possible to define a reliable *relevance* measure such that the speaker may simply maximise the relevance of his answers for optimally supporting the hearer. When solving a support problem the speaker has to take the hearer’s response to his choice of signal into account. Hence, in view of our previous discussion of IBR models, this shows that there is no reliable method of solving a support problem which involves fewer steps of reasoning about each other than backward induction. Support problems incorporate Grice’s *Cooperative Principle*, his maxim of *Quality*, and a method for finding optimal strategies which replaces Grice’s maxims of *Quantity* and *Relevance*. For now, we ignore the maxim of *Manner*.

A decision problem consists of a set Ω of the possible states of the world, the decision maker’s expectations about the world, a set of actions \mathcal{A} he can choose from, and his preferences regarding their outcomes. We always assume that Ω is finite. We represent an agent’s expectations about the world by a probability distribution over Ω , i.e. a real valued function $P : \Omega \rightarrow \mathbb{R}$ with the following properties: (1) $P(v) \geq 0$ for all $v \in \Omega$ and (2) $\sum_{v \in \Omega} P(v) = 1$. For sets $F \subseteq \Omega$ it is $P(F) = \sum_{v \in F} P(v)$. The pair (Ω, P) is called a finite *probability space*. An agent’s preferences regarding outcomes of actions are represented by a real valued function over world–action pairs. We collect these elements in the following structure:

Definition 4.1 A decision problem is a triple $\langle (\Omega, P), \mathcal{A}, u \rangle$ such that (Ω, P) is a finite probability space, \mathcal{A} a finite, non–empty set and $u : \Omega \times \mathcal{A} \rightarrow \mathbb{R}$

a function. \mathcal{A} is called the action set, and its elements actions; u is called a payoff or utility function.

In the following, a decision problem $\langle(\Omega, P), \mathcal{A}, u\rangle$ represents the hearer's situation before receiving information from an answering expert. We will assume that this problem is common knowledge. How to find a solution to a decision problem? It is standard to assume that rational agents try to maximise their expected utilities. In Section 2, we used the symbol \mathcal{E} to denote the expected utility. As in the present section probabilities are assumed to be subjective probabilities, we use different notation in order to distinguish subjective expected utilities from expected utilities defined from objective frequencies. Hence, we write for the (subjective) *expected utility* of action $a \in \mathcal{A}$ in decision problem $\langle(\Omega, P), \mathcal{A}, u\rangle$:

$$EU(a) = \sum_{v \in \Omega} P(v) \times u(v, a). \quad (4.27)$$

The expected utility of actions may change if the decision maker learns new information. To determine this change of expected utility, we first have to know how learning new information affects the hearer's beliefs. In probability theory the result of learning a proposition F is modelled by *conditional probabilities*. Let H be any proposition and F the newly learned proposition. Then, the probability of H given F , written $P(H|F)$, is defined as

$$P(H|F) := P(H \cap F) / P(F) \text{ for } P(F) \neq 0. \quad (4.28)$$

In terms of this conditional probability function, the *expected utility after learning F* is defined as

$$EU(a|F) = \sum_{v \in \Omega} P(v|F) \times u(v, a). \quad (4.29)$$

H will choose the action which maximises his expected utilities after learning F , i.e. he will only choose actions a for which $EU(a|F)$ is maximal. We assume that H 's decision does not depend on what he believes that the answering speaker believes. We denote the set of actions with maximal expected utility by $\mathcal{B}(F)$, i.e.

$$\mathcal{B}(F) := \{a \in \mathcal{A} \mid \forall b \in \mathcal{A} \ EU_H(b|F) \leq EU_H(a|F)\}. \quad (4.30)$$

The decision problem represents the hearer's situation. In order to get a model of the questioning and answering situation, we have to add a representation of the answering speaker's information state. We identify it with a (subjective) probability distribution P_S that represents his expectations about the world. We make a number of assumptions in order to match the definition of support problems to our previous definition of signalling games. First, we assume that

the hearer's expectations are common knowledge. Second, we assume that there exists a common prior from which both the speaker's and the hearer's information state can be derived by a Bayesian update. This entails that the speakers and the hearer's expectations cannot contradict each other. Third, we assume that the speaker does not directly choose propositions but linguistic *forms* or *signals* which have a predefined semantics. Furthermore, we assume that the forms $F \in \mathcal{F}$ come with positive costs. This leads to the following definition of *interpreted* support problems:

Definition 4.2 A tuple $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ is an interpreted support problem if:

1. (Ω, P_S) is a finite probability space and $\langle (\Omega, P_H), \mathcal{A}, u \rangle$ a decision problem;
2. there exists a probability distribution P on Ω , and sets $K_S \subseteq K_H \subseteq \Omega$ for which $P_S(X) = P(X|K_S)$ and $P_H(X) = P(X|K_H)$;
3. $\llbracket \cdot \rrbracket : \mathcal{F} \rightarrow \mathcal{P}(\Omega)$ is an interpretation function for the elements $F \in \mathcal{F}$. We assume that

$$\forall X \subseteq \Omega \exists F \in \mathcal{F} \llbracket F \rrbracket = X; \quad (4.31)$$

4. $u : \Omega \times \mathcal{A} \rightarrow \mathbb{R}$ is a utility measure and c a cost function that maps forms $F \in \mathcal{F}$ to positive real number.

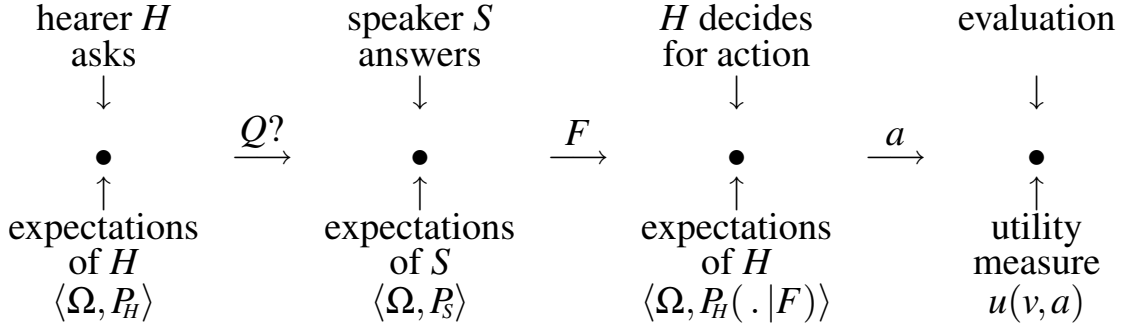
The second condition says that P_S and P_H are derived from a common prior P by a Bayesian update. It entails:

$$\forall X \subseteq \Omega P_S(X) = P_H(X|K_S). \quad (4.32)$$

This condition allows us to identify the *common ground* in conversation with the addressee's expectations about the domain Ω , i.e. with P_H . The speaker knows the addressee's information state and is at least as well informed about Ω . Hence, the assumption is a probabilistic equivalent to the assumption about common ground that implicitly underlies dynamic semantics (Groenendijk and Stockhof, 1991). Furthermore, condition (4.32) implies that the speaker's beliefs cannot contradict the hearer's expectations, i.e. for $X \subseteq \Omega$: $P_S(X) = 1 \Rightarrow P_H(X) > 0$.

In order to simplify notation, we will often write F instead of $\llbracket F \rrbracket$. Hence, F may denote a proposition or a linguistic form, depending on context.

Our next goal is to introduce a principle for solving support problems, i.e. for finding the speaker's and hearer's strategies which lead to optimal outcomes. The speaker S 's task is to provide information that is optimally suited to support H in his decision problem. Hence, we find two successive decision problems, in which the first problem is S 's problem to choose an answers. The utility of the answer depends on how it influences H 's final choice:



We assume that S is fully cooperative and wants to maximise H 's final success; i.e. S 's payoff, is identical with H 's. This is our representation of Grice's *Cooperative Principle*. S has to choose an answer that induces H to choose an action that maximises their common payoff. In general, there may exist several equally optimal actions $a \in \mathcal{B}(F)$ which H may choose. Hence, the expected utility of an answer depends on the probability with which H will choose the different actions. We can assume that this probability is given by a probability measure $h(\cdot|F)$ on \mathcal{A} . Then, the expected utility of an answer F is defined by:

$$EU_s(F) := \sum_{a \in \mathcal{B}(F)} h(a|F) \times EU_s(a). \quad (4.33)$$

We add here a further Gricean maxim, the *Maxim of Quality*. We call an answer F *admissible* if $P_s(F) = 1$. The Maxim of Quality is represented by the assumption that the speaker S does only give admissible answers. This means that he believes them to be *true*. For an interpreted support problem $\sigma = \langle \Omega, P_s, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ we set:

$$\text{Adm}_\sigma := \{F \subseteq \Omega \mid P_s(F) = 1\} \quad (4.34)$$

Hence, the set of optimal answers in σ is given by:

$$\text{Op}_\sigma := \{F \in \text{Adm}_\sigma \mid \forall B \in \text{Adm}_\sigma EU_s(B) \leq EU_s(F)\}. \quad (4.35)$$

We write Op_σ^h if we want to make the dependency of Op on h explicit. Op_σ is the set of *optimal answers* for the support problem σ . Condition (4.31), it follows that all propositions $A \subseteq \Omega$ can be expressed. Hence, we can think of Op_σ as a subset of $\mathcal{P}(\Omega)$ or as a subset of \mathcal{F} .

The *behaviour* of interlocutors can be modelled by *strategies*. A strategy is a function which tells us for each information state of an agent which actions he may choose. It is not necessary that a strategy picks out a unique action for each information state. A *mixed* strategy is a strategy which chooses actions with certain probabilities. The hearer strategy $h(\cdot|F)$ is an example of a mixed strategy. We define a (mixed) strategy pair for an interpreted support problem σ to be a pair (s, h) such that s is a probability distribution over \mathcal{F} and $h(\cdot|F)$ a probability distribution over \mathcal{A} .

We may call a strategy pair (s, h) a *solution* to σ iff $h(\cdot|F)$ is a probability distribution over $\mathcal{B}(F)$, and s a probability distribution over Op_σ^h . In general, the solution to a support problem is not uniquely defined. Therefore, we introduce the notion of the *canonical* solution.

Definition 4.3 Let $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ be a given interpreted support problem. The canonical solution to σ is a pair (S, H) of mixed strategies which satisfy:

$$S(F) = \begin{cases} |\text{Op}_\sigma|^{-1}, & F \in \text{Op}_\sigma \\ 0 & \text{otherwise} \end{cases}, \quad H(a|F) = \begin{cases} |\mathcal{B}(F)|^{-1}, & a \in \mathcal{B}(F) \\ 0 & \text{otherwise} \end{cases}. \quad (4.36)$$

We write $S(\cdot|\sigma)$ if S is a function that maps each $\sigma \in \mathcal{S}$ to the speaker's part of the canonical solution, and $H(\cdot|D_\sigma)$ if H is a function that maps the associated decision problem D_σ to the hearer's part of the canonical solution. From now on, we will always assume that speaker and hearer follow the canonical strategies $S(\cdot|\sigma)$ and $H(\cdot|D_\sigma)$. We make this assumption because it is convenient to have a unique solution to a support problem; the only property that we really need in the following proofs is that $H(a|F) > 0 \Leftrightarrow a \in \mathcal{B}(F)$ and $S(F|\sigma) > 0 \Leftrightarrow F \in \text{Op}_\sigma$.

The speaker may always answer everything he knows, i.e. he may answer $K_S := \{v \in \Omega \mid P_S(v) > 0\}$. Condition (4.32) trivially entails that $\mathcal{B}(K_S) = \{a \in \mathcal{A} \mid \forall b \in \mathcal{A} \text{ } EU_S(b) \leq EU_S(a)\}$. If speaker and hearer follow the canonical solution, and if we ignore the different costs of answers, then:

$$\text{Op}_\sigma = \{F \in \text{Adm}_\sigma \mid \mathcal{B}(F) \subseteq \mathcal{B}(K_S)\}. \quad (4.37)$$

In order to show (4.37), let $F \in \text{Adm}$ and $\alpha := \max\{EU_S(a) \mid a \in \mathcal{A}\}$. For $a \in \mathcal{B}(F) \setminus \mathcal{B}(K_S)$ it holds by definition that $EU_S(a) < \alpha$ and $H(a|F) > 0$. $EU_S(F)$ is the sum of all $H(a|F) \times EU_S(a)$. If $\mathcal{B}(F) \not\subseteq \mathcal{B}(K_S)$, then this sum divides into the sum over all $a \in \mathcal{B}(F) \setminus \mathcal{B}(K_S)$ and all $a \in \mathcal{B}(F) \cap \mathcal{B}(K_S)$. Hence, $EU_S(F) < \alpha$, and therefore $F \notin \text{Op}_\sigma$.

If $\mathcal{B}(F) \not\subseteq \mathcal{B}(K_S)$, then the speaker knows that answering F would induce the addressee to choose a sub-optimal action with positive probability. In this sense, we can call an answer F *misleading* if $\mathcal{B}(F) \not\subseteq \mathcal{B}(K_S)$; then, (4.37) implies that Op_σ is the set of all non-misleading answers.

4.3 Signalling Games and the Optimal Answer Model

We first recall the definition of signalling games from the previous sections. A *signalling game* is a tuple $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ for which: (1) Ω and Θ are non-empty finite sets; (2) $P(\cdot)$ is a probability distribution over Ω ; (3) $p(\cdot|v)$ is a probability distribution over Θ for every $v \in \Omega$; (4) \mathcal{F} and \mathcal{A} are respectively

the speaker's and hearer's action sets; and (5) $u : \Omega \times \Theta \times \mathcal{F} \times \mathcal{A} \rightarrow \mathbb{R}$ is a shared utility function. We also assumed that $u(v, \theta, F, a)$ can be decomposed into $u(v, a) - c(F)$ for some positive value $c(F)$.

We first discuss the consequences of interpreting the probabilities for signalling games as objective frequencies and that for support problems as subjective probabilities.

If \mathcal{S} is a set of support problems with identical decision problems, we can construct a corresponding signalling game. As it is assumed that the speaker knows the full support problem, we can identify \mathcal{S} with the set of speaker's types Θ . The action sets and the utility function of the signalling game are just the same as that of the support problems. As the decision problems of the support problems in \mathcal{S} are identical, this poses no problem. The only non-trivial correspondence is that of the probabilities.

As mentioned before, we regard the probabilities P and p of the signalling game as objective frequencies. Under this interpretation, Lemma 2.3 states the objective conditions for optimal signalling strategies. If we interpret P_S^σ and P_H^σ as the agents' representations for these objective probabilities, then P_S must be identical to μ_S , and P_H to P .³ (4.32) then entails that $P_H(v|K_S) = \mu_S(v|\sigma)$. It holds $P_H(v|K_S) = \mu_S(v|\sigma)$ iff $P(v)/P(K_S) = P(v) p(\sigma|v)/\mu(\sigma)$ iff $p(\sigma|v) = \mu(\sigma)/P(K_S)$. The last term does not depend on v , hence, it follows that (4.32) entails that $p(\sigma|v)$ must be the same for all $v \in K_S$.

In (Benz and van Rooij, 2007), we identified P_S with $P(\cdot | K_S)$, and P_H with P . Then (4.32) trivially holds. p was considered to be a representation of the hearer's subjective expectations about the speaker's types. In order to distinguish the hearer's subjective probabilities about the speaker's type from the objective frequencies, we write p_H for the former, and keep p for the latter. Subjective probabilities per se have no causal influence on the objective probabilities. Hence, p_H is logically independent from P and p . Under this interpretation, it can be shown that the strategy pair (S, H) defined by the canonical solutions to the support problems (4.36) is optimal for all possible p_H . This result follows from Lemma 2.3 if we assume that the objective frequencies represented by p in the signalling game again satisfy $p(\sigma|v) = \mu(\sigma)/P(K_S)$. Then, whatever the subjective expectations of the hearer about the speaker's types are, the canonical strategy will satisfy (2.9), and hence be optimal in the sense that there is no other strategy pair with higher expected utility.

In this paper, we go one step further and completely separate the subjective cognitive level from the objective level. Hence, we interpret the probabilities P_S and P_H in the support problems as subjective probabilities which are logically independent of the frequencies P and p of the underlying signalling game. As P_S and P_H are subjective, they don't change the objective information

³The probabilities μ_S and μ have been defined in (2.2) and (2.3).

available to S and H . Hence, we can freely assign these probabilities to the interlocutors without changing the signalling game on the objective level. Subjective probabilities determine the speaker's and the hearer's strategies. These strategies are the only connection between the cognitive and the realistic level.

What is the advantage of separating the cognitive and the objective level? There are two issues involved: the *epistemic* issue of the recognisability of objective frequencies, and the issue of *bounded rationality*. For the epistemic issue, the objective frequencies are largely unknown to the interlocutors. The speaker may learn his type θ e.g. by direct observation, by an inductive inference, by hear-say, or from a conversation with someone else. Hence, there are so many and so varied sources for the acquisition of belief type θ that it is not to be expected that the hearer or the speaker can provide any justified estimate of $p(\theta|v)$. In this respect, conversation can be characterised as a game of *complete uncertainty*. Even though, we can assign rationally justified subjective probabilities which describe the agent's behaviour on the cognitive level. This move allows us to treat communication as a game under *risk*. For the issue of bounded rationality, it doesn't deem us a realistic assumption that interlocutors do an online calculation of their conditional probabilities μ_S and μ_H defined in (2.2). The established solution concept for signalling games is that of a perfect Bayesian equilibrium. Hence, even if we could assume that the interlocutors know the objective frequencies P and p , the complexity of calculating the Bayesian perfect equilibria would make the resulting model cognitively implausible. By separating the cognitive and the objective level of reality, we can justify simpler solutions to the coordination problem, and at the same time explain their objective success.

What is our approach to the problem of bounded rationality? If we want to show that a strategy pair (S, H) is a successful solution to a signalling game, we have to show that it is a Perfect Bayesian equilibrium in the objective sense. We will even show that the strategies established on the cognitive level are such that they Pareto dominate all other solutions. Hence, our strategy for solving the problem of bounded rationality is to search for the simplest solution on the cognitive level that can guarantee objective success. As the discussion of relevance scale approaches in (Benz, 2006, 2007) shows, the interlocutors have to solve a game theoretic problem, i.e. it is not possible to guarantee objective communicative success by simply applying decision theoretically defined solutions on the cognitive level. Signalling games are sequential games. The simplest solution to a sequential game is that found by backward induction. Hence, the optimal answer model claims that the most simple solution concept for sequential games is already successful. Moreover, it involves that the hearer does not need to take his expectations p_H about the speaker's types θ into account. This leads to our main criterion of simplicity: we assume that a method for finding

a solution (S, H) is the simpler the less reasoning about each other is involved in it. In terms of the IBR model, this means that a R_0 – S_1 reasoning sequence is sufficient for finding reliable stable equilibria.

In order to decide whether the canonical strategy determined by a set of support problems is a Pareto optimal equilibrium for the related signalling game, the logical relation between the objective frequencies of signalling games and the subjective probabilities of sets of support problems play a central role. We consider the following relations:

Definition 4.4 *Let \mathcal{S} be a set of interpreted support problems. Let's assume that the support problems $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ may only differ with respect to P_S^σ . Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be any basic signalling game for which $\Theta = \mathcal{S}$ and $\mu_\Theta(\sigma) = \sum_v P(v) p(\sigma|v) > 0$ for all $\sigma \in \mathcal{S}$. We call the speaker's probability P_S^σ :*

1. fully reliable if $P_S^\sigma = \mu_s(\cdot | \sigma)$.
2. reliable if $\forall v \in \Omega (\mu_s(v | \sigma) > 0 \Leftrightarrow P_S^\sigma(v) > 0)$.
3. truth preserving if $\forall v \in \Omega (\mu_s(v | \sigma) > 0 \Rightarrow P_S^\sigma(v) > 0)$.

We say that:

4. \mathcal{G} supports \mathcal{S} iff all P_S^σ are reliable;
5. \mathcal{G} fully supports \mathcal{S} iff all P_S^σ are fully reliable;
6. \mathcal{G} weakly supports \mathcal{S} iff all P_S^σ are truth preserving.

Full reliability is stronger than reliability, and reliability is stronger than truth preservingness. If P_S is truth preserving then all believes of S are true in the sense that $P_S^\sigma(F) = 1$ implies that the true state of the world must be an element of F . This follows from $P(v) = 0 \Rightarrow \mu_s(v | \sigma) = P(v) p(\sigma|v) = 0$.

Furthermore, we introduce two conventions: (1) If the support problem does not specify a set of utterances \mathcal{F} or costs of signals, then we assume that for supporting signalling games it holds that $\mathcal{F} = \mathcal{P}(\Omega)$, and that $u(v, \theta, F, a)$ does only depend on v and a . (2) We also use the terminology of Def. 4.4 if Θ and \mathcal{S} can only be identified with each other by a bijective map. In this case, we write θ_σ and σ_θ for the speaker type and the support problem which have been identified with each other.

The following two lemmas provide the justification for the optimal answer approach. The first one tells us that the canonical solution to a set of support problems is a Pareto Nash equilibrium for all fully supporting signalling games. The second lemma strengthens this result for support problems with expert speaker. In this case, the canonical solution is a Pareto Nash equilibrium to all weakly supporting signalling games.

Lemma 4.5 *Let \mathcal{S} be a set of interpreted support problems. Let's assume that the support problems $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ may only differ with respect to P_S^σ . Let (S, H) be the canonical solution to \mathcal{S} . Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be any basic signalling game which fully supports \mathcal{S} , i.e. $\Theta = \mathcal{S}$ and the speaker's probabilities P_S^σ are fully reliable. Then (S, H) is a Pareto Nash equilibrium of \mathcal{G} .*

Proof: The lemma follows if we can show that the canonical solution satisfies (2.9) for all $F \in \mathcal{F}$. Hence, let F be given, and σ be such that $\exists v P(v) p(\sigma|v) > 0$. By definition, $S(F|\sigma) > 0$ iff $F \in \text{Op}_\sigma$; hence, it follows from (4.37) and the definition of the canonical hearer strategy that $H(a|F) > 0$ entails $a \in \mathcal{B}(K_S^\sigma)$ with $K_S^\sigma = \{v \in \Omega \mid P_S^\sigma(v) > 0\}$. As P_S is fully reliable, it follows that $\mathcal{B}(K_S^\sigma) = \mathcal{B}(\sigma)$, and therefore that $H(a|F) > 0 \Rightarrow a \in \mathcal{B}(\sigma)$. Hence, $S(F|\sigma) > 0 \Rightarrow H(\mathcal{B}(\sigma)|F) = 1$. ■

For support problems with expert speakers, we arrive at a stronger result:

Lemma 4.6 *Let \mathcal{S} be a set of interpreted support problems. Let's assume that the support problems $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ may only differ with respect to P_S^σ . Let us further assume that the speaker is an expert, i.e.*

$$\forall \sigma \in \mathcal{S} \exists a \in \mathcal{A} P_S^\sigma(O(a)) = 1.$$

Let (S, H) be the canonical solution to \mathcal{S} . Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be any signalling game which weakly supports S . Then (S, H) is a Pareto Nash equilibrium of \mathcal{G} .

Proof: That the speaker is an expert entails that $\mathcal{B}(K_S^\sigma) = \{a \in \mathcal{A} \mid P_S^\sigma(O(a)) = 1\}$. As $\mu_S(v|\sigma) > 0 \Rightarrow P_S^\sigma(v) > 0$, it follows that $\mathcal{B}(K_S^\sigma) \subseteq \mathcal{B}(\sigma)$. Hence, the claim follows as in the proof of Lemma 4.5. ■

It is an obvious question, how to construct a signalling game \mathcal{G} for a given set of support problems \mathcal{S} so that \mathcal{G} is fully supporting \mathcal{S} . The answer will be provided by the next lemma. Finally, we will also address the question how and when we can construct a set \mathcal{S} of support problems for a given signalling game \mathcal{G} such that \mathcal{G} supports \mathcal{S} .

Lemma 4.7 *Let \mathcal{S} be a set of interpreted support problems. Let's assume that the support problems $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ may only differ with respect to P_S^σ . Let μ be any probability measure on \mathcal{S} for which $\mu(\sigma) > 0$ for all $\sigma \in \mathcal{S}$. Then let $v(v, \sigma) := \mu(\sigma) P_S^\sigma(v)$, $P(v) := \sum_{\sigma} v(v, \sigma)$, and $p(\sigma|v) := v(v, \sigma)/P(v)$. Then v is a probability measure on $\Omega \times \mathcal{S}$, and $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ is fully supporting \mathcal{S} .*

Proof: As $\sum_{v, \sigma} \mu(\sigma) P_S^\sigma(v) = \sum_{\sigma} \mu(\sigma) \sum_v P_S^\sigma(v) = 1$, v is a probability measure on $\Omega \times \mathcal{S}$. That \mathcal{G} supports \mathcal{S} follows from $\mu_\Theta(\sigma) = \sum_w P(w) p(\sigma|w) =$

$\sum_w v(w, \sigma) = \mu(\sigma) \sum_w P_S^\sigma(w) = \mu(\sigma)$; hence $\mu_\Theta(\sigma) > 0$ for all $\sigma \in \mathcal{S}$. Finally, $\mu_S(v|\sigma) = \frac{P(v)p(\sigma|v)}{\sum_w P(w)p(\sigma|w)} = \frac{v(v,\sigma)}{\sum_w v(w,\sigma)} = \frac{P_S^\sigma(v)}{\sum_w P_S^\sigma(w)} = P_S^\sigma(v)$. Hence, $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ is fully supporting \mathcal{S} . ■

The inverse construction is not always possible. We already have seen that (4.32) entails that, for signalling games which fully support a set of support problems, $p(\theta|v)$ must be the same for all $v \in K_S$. Hence, there cannot be for every signalling game a set of support problems which is fully supported by it. If \mathcal{G} is such that $p(\theta|v)$ is the same for all $v \in K_S^\theta := \{v \in \Omega \mid \mu(v|\theta) > 0\}$, then we can set $P_S^\theta(v) := P(v|K_S^\theta)$ and $P_H^\theta(v) := P(v|K_H^\theta)$ with $K_H^\theta := \{v \in \Omega \mid P(v) > 0\}$. Then K_H^θ and P_H^θ do not depend on θ , and we find $\mu(v|\theta) = P(v)p(\theta|v)/\sum_w (P(w)p(\theta|w)) = P(v)/P(K_S^\theta) = P(v|K_S^\theta) = P_H(v|K_S^\theta) = P_S^\theta(v)$.

For the general case, we either have to give up (4.32) or full reliability. If we decide to give up (4.32), then we can set $P_S^\theta = \mu(v|\theta)$ and e.g. $P_H(v) = P(v)$, and arrive for each θ at a support problem with fully reliable speaker expectations. If we decide to give up full reliability, then we can set $P_S^\theta(v) = P(v|K_S^\theta)$ and $P_H = P$, and arrive for each θ at a reliable support problem which satisfies (4.32). In either case, P_H does not depend on θ . Hence, the support problems in the constructed set \mathcal{S} do only differ with respect to P_S .

We summarise the result:

Lemma 4.8 *Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a given signalling game. For $\theta \in \Theta$, let $K_S^\theta := \{v \in \Omega \mid \mu(v|\theta) > 0\}$, $P_S^\theta(v) := P(v|K_S^\theta)$, and $P_H^\theta := P$. Let σ_θ be the resulting support problem. Then, the P_S^θ are reliable, and it holds:*

1. *the support problems σ_θ satisfy (4.32): $P_S^{\sigma_\theta} = P(v|K_S^\theta) = P_H(v|K_S^\theta)$.*
2. *If, in addition, $p(\theta|v)$ is the same for all $v \in K_S^\theta$, then the support problems σ_θ are also fully reliable, i.e. $P_S^{\sigma_\theta} = \mu(\cdot|\theta)$.*

Support problems which do not satisfy (4.32) were considered in (Benz, 2006).

5 Implicatures

In this section, we apply the ideas of Section 3 to signalling games and prove more explicit characterisations of implicatures. We assume throughout that a fixed signalling game $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ together with a strategy pair (S, H) is given. As explained in Section 3, $\langle \mathcal{G}, S, H \rangle$ defines a linear causal Bayesian network $(\mathcal{X}_i, p_i)_{i=0,\dots,3}$ if we identify \mathcal{X}_0 with Ω , \mathcal{X}_1 with Θ , \mathcal{X}_2 with \mathcal{F} , and \mathcal{X}_3 with \mathcal{A} ; accordingly, we set $p_0 = P$, $p_1 = p$, $p_2 = S$, and $p_3 = H$. In this section, we write $\mu(\theta)$ and $\mu(F)$ for the marginal probabilities $\mu_1(\theta)$ and $\mu_2(F)$, and $\mu(\theta|F)$ for the conditional marginal probability $\mu_{1|2}(\theta|F)$.⁴

⁴The definitions of these probability distributions in the form of explicit sums can be found in (2.3) and (6.52).

We write, by a small mis-use of logical notation, $\langle \mathcal{G}, S, H \rangle \models F +> R$ if the utterance of F implicates R . In Lemma 3.4, we have shown that for any set $Y \subseteq \mathcal{F}$ and $X \subseteq \mathcal{X}_i$ it holds that:

$$\langle \mathcal{G}, S, H \rangle \models Y +> X \iff \mu_{i|\mathcal{F}}(X|Y) = 1 \quad (5.38)$$

In traditional theories of implicatures, it is assumed that an implicature provides information about the world or the speaker's information state in addition to the literally communicated information. Therefore, we are now concentrating on the cases $\mathcal{X}_i = \Omega$ and $\mathcal{X}_i = \Theta$; i.e. we are looking for a characterisation of implicatures *about the world* and the *speaker's state*. For $F \subseteq \mathcal{F}$ with $\mu_{\mathcal{F}}(F) > 0$, and $R_0 \subseteq \Omega$ or $R_1 \subseteq \Theta$, the criterion in (5.38) reads as:

$$\langle \mathcal{G}, S, H \rangle \models F +> R_i \iff \mu(R_i|F) = 1. \quad (5.39)$$

By definition, $\mu(R_i|F) = 1$ is equivalent to

$$\frac{\mu(\pi_i^{-1}[R_i] \cap \pi_{\mathcal{F}}^{-1}[F])}{\mu(\pi_{\mathcal{F}}^{-1}[F])} = 1. \quad (5.40)$$

We first consider R_1 , which is a subset of Θ . Then (5.40) is equivalent to $\{\theta \in R_1 \mid \mu_{\Theta}(\theta) > 0 \wedge S(F|\theta) > 0\} \supseteq \{\theta \in \Theta \mid \mu_{\Theta}(\theta) > 0 \wedge S(F|\theta) > 0\}$. If $\mu_{\Theta}(\theta) > 0$ for all $\theta \in \Theta$, then this formula is again equivalent to $\forall \theta : S(F|\theta) > 0 \Rightarrow \theta \in R_1$.

We now turn to the implicatures about the state of the world, i.e. to R_0 , which is a subset of Ω . Then (5.40) is equivalent to $\{v \in R_0 \mid P(v) > 0 \wedge \exists \theta (p(\theta|v) > 0 \wedge S(F|\theta) > 0)\} \supseteq \{v \in \Omega \mid P(v) > 0 \wedge \exists \theta (p(\theta|v) > 0 \wedge S(F|\theta) > 0)\}$. If $P(v) > 0$ for all $v \in \Omega$, then this formula is again equivalent to $\forall v : \mu(F|v) > 0 \Rightarrow v \in R_0$.

We summarise this result in the following proposition:

Proposition 5.1 *Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a signalling game and (S, H) a strategy pair. Let $F \subseteq \mathcal{F}$ with $\mu_{\mathcal{F}}(F) > 0$. Then it holds:*

1. *If $R \subseteq \Theta$, and if for all $\theta \in \Theta$ $\mu_{\Theta}(\theta) > 0$, then*

$$\langle \mathcal{G}, S, H \rangle \models F +> R \iff \forall \theta : S(F|\theta) > 0 \Rightarrow \theta \in R. \quad (5.41)$$

2. *If $R \subseteq \Omega$, and if for all $v \in \Omega$ $P(v) > 0$, then*

$$\langle \mathcal{G}, S, H \rangle \models F +> R \iff \forall v : \mu(F|v) > 0 \Rightarrow v \in R. \quad (5.42)$$

Note that the implicatures are completely independent of the meaning of the signals in \mathcal{F} . Hence, they are also defined for situations in which the signals have no pre-defined semantic meaning. The implicature of a signal coincides with Lewis notion of *indicated meaning* (2002). Lewis *defined* the semantic

meaning of signals as their indicated meaning. In this way, he could explain how the semantics of signals can emerge from a convention about their use. If we assume that a semantics is already established, then the indicated meaning may exceed this pre-defined semantic meaning. This additional information is commonly called an implicature. Our definition in (5.39) differs from common usage of the word *implicature* in so far as the literal meaning of a signal, if defined, is subsumed by implicated meaning. We can define a stronger notion of implicature which is more in accordance with the common usage. According to this notion, an utterance of F implicates R only if R does not already follow from the communicated semantic meaning of F . We only introduce this notion in order to show that an equivalent to the common notion of implicature can easily be derived from our definition; but the concept of proper implicatures will not be used anywhere in this paper.

Definition 5.2 (Proper Implicatures) *Let \mathcal{S} be a set of interpreted support problems $\langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ which may only differ with respect to P_S . Let (S, H) be a strategy pair for \mathcal{S} . For $R \subseteq \Omega$, $F \in \mathcal{F}$, and $\llbracket F \rrbracket^* := \{v \in \llbracket F \rrbracket \mid P_H(v) > 0\}$, we say that the utterance of F properly implicates that R in $\langle \mathcal{S}, S, H \rangle$ iff $\langle \mathcal{S}, S, H \rangle \models F +> R \& \llbracket F \rrbracket^* \setminus R \neq \emptyset$.*

We now turn our attention to support problems. In (Benz, 2008), the implicatures $R \subseteq \Omega$ of a sentence F in a given set of support problems \mathcal{S} were defined by $\langle \mathcal{S}, S, H \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (F \in \text{Op}_\sigma \Rightarrow P_S^\sigma(R) = 1)$. We now show that:

Lemma 5.3 *Let \mathcal{S} be a set of support problems $\sigma = \langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ which only differ with respect to P_S^σ . Let \mathcal{G}_0 and \mathcal{G}_1 both be signalling games which support \mathcal{S} . Let (S, H) be a pair of signalling strategies for \mathcal{G}_0 and \mathcal{G}_1 . Then, it holds:*

1. $\mu^{\langle \mathcal{G}_0, S, H \rangle}(F) > 0$ iff $\mu^{\langle \mathcal{G}_1, S, H \rangle}(F) > 0$.
2. If $\mu^{\langle \mathcal{G}_i, S, H \rangle}(F) > 0$ and $R \subseteq \Omega$, then it holds:

$$\langle \mathcal{G}_i, S, H \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (S(F|\sigma) > 0 \Rightarrow P_S^\sigma(R) = 1). \quad (5.43)$$

3. If $\mu^{\langle \mathcal{G}_i, S, H \rangle}(F) > 0$ and $R \subseteq \Omega$, then it holds:

$$\langle \mathcal{G}_0, S, H \rangle \models F +> R \iff \langle \mathcal{G}_1, S, H \rangle \models F +> R. \quad (5.44)$$

Proof: That the \mathcal{G}_i support \mathcal{S} implies, by Def. 4.4, that for all $v \in \Omega$: $P_S^\sigma(v) > 0 \iff P^{\mathcal{G}_i}(v) p^{\mathcal{G}_i}(\sigma|v) > 0$. By definition of $\mu(F)$ in (2.3), $\mu^{\langle \mathcal{G}_i, S, H \rangle}(F) > 0$ iff $\sum_{v, \sigma} P^{\mathcal{G}_i}(v) p^{\mathcal{G}_i}(\sigma|v) S(F|\sigma) > 0$. As the \mathcal{G}_i support \mathcal{S} , the latter is equivalent to $\sum_{v, \sigma} P_S^\sigma(v) S(F|\sigma) > 0$. From this, the first claim follows immediately.

Let us now only consider \mathcal{G}_0 . Let $\mu(v, \sigma, F) := P(v) p(\sigma|v) S(F|\sigma)$. Then, with (5.42) we find

$$\begin{aligned} \langle \mathcal{G}_0, S, H \rangle \models F +> R &\Leftrightarrow \forall v : \mu(F|v) > 0 \Rightarrow v \in R \\ &\Leftrightarrow \{ \langle v, \sigma, F \rangle \mid \mu(v, \sigma, F) > 0 \} \subseteq \{ \langle v, \sigma, F \rangle \mid \mu(v, \sigma, F) > 0 \wedge v \in R \}. \end{aligned} \quad (5.45)$$

As all P_s^σ are reliable, the last condition is equivalent to $\{ \langle v, \sigma, F \rangle \mid P_s^\sigma(v) > 0 \wedge S(F|\sigma) > 0 \} \subseteq \{ \langle v, \sigma, F \rangle \mid P_s^\sigma(v) > 0 \wedge S(F|\sigma) > 0 \wedge v \in R \}$. This is again equivalent to $\forall v, \sigma (P_s^\sigma(v) > 0 \wedge S(F|\sigma) > 0 \Rightarrow v \in R)$, which finally is equivalent to $\forall \sigma (S(F|\sigma) > 0 \Rightarrow P_s^\sigma(R) = 1)$. This proves the second claim. The third claim follows immediately from the second. ■

With (5.44), we can define:

Definition 5.4 Let \mathcal{S} be a set of support problems σ which only differ with respect to P_s^σ . Let $\text{Supp}(\mathcal{S})$ be the set of all signalling games \mathcal{G} which support \mathcal{S} . Let (S, H) be any strategy pair for \mathcal{S} , and let $F \in \mathcal{F}$ be such that $\exists \sigma \in \mathcal{S} S(F|\sigma) > 0$. Then we set for $R \subseteq \Omega$:

$$\langle \mathcal{S}, S, H \rangle \models F +> R \iff \forall \mathcal{G} \in \text{Supp}(\mathcal{S}) \langle \mathcal{G}, S, H \rangle \models F +> R. \quad (5.46)$$

Note that by Lemma 4.7 $\text{Supp}(\mathcal{S})$ is never empty. If (S, H) is the *canonical solution* to \mathcal{S} , we arrive with (5.43) at:

$$\langle \mathcal{S}, S, H \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (F \in \text{Op}_\sigma \Rightarrow P_s^\sigma(R) = 1). \quad (5.47)$$

Starting from (5.47), we can derive criteria for special but frequent situations. The remainder of the section presents some results from (Benz, 2008).

First, we note that, as the hearer has to check all support problems in \mathcal{S} , we arrive at the more implicatures the smaller \mathcal{S} becomes. If $\mathcal{S} = \{\sigma\}$ and $F \in \text{Op}_\sigma$, then F will implicate everything the speaker knows. The other extreme is the case in which answers implicate only what they logically entail. We show in Proposition 5.7 that this case can occur.

We are interested in cases in which the speaker is a real expert. If he is an expert, then we can show that there is a very simple criterion for calculating implicatures. We can call the speaker an expert if he knows the actual world; but we will see that a weaker condition is sufficient for our purposes. To make precise what we mean by expert, we introduce another important notion, the set $O(a)$ of all worlds in which an action a is optimal:

$$O(a) := \{w \in \Omega \mid \forall b \in \mathcal{A} u(w, a) \geq u(w, b)\}. \quad (5.48)$$

We say that the answering person is an expert for a decision problem if there is an action which is an optimal action in all his epistemically possible worlds. We represent this information in \mathcal{S} :

Definition 5.5 (Expert) Let \mathcal{S} be a set of support problems with joint decision problem $\langle(\Omega, P_H), \mathcal{A}, u\rangle$. Then we call S an expert in a support problem σ if $\exists a \in \mathcal{A} P_s^\sigma(O(a)) = 1$. He is an expert in \mathcal{S} , if he is an expert in every $\sigma \in \mathcal{S}$.

This leads us to the following criterion for implicatures:

Lemma 5.6 Let \mathcal{S} be a set of support problems with joint decision problem $\langle(\Omega, P_H), \mathcal{A}, u\rangle$, and (S, H) its canonical solution. Assume furthermore that E is an expert for every $\sigma \in \mathcal{S}$ and that $\forall v \in \Omega \exists \sigma \in \mathcal{S} P_s^\sigma(v) = 1$. Let $\sigma \in \mathcal{S}$ and $F, R \subseteq \Omega$ be two propositions with $F \in \text{Op}_\sigma$. Then, with $F^* := \{v \in \Omega \mid P_H(v) > 0\}$, it holds that:

$$\langle \mathcal{S}, S, H \rangle \models F +> R \text{ iff } F^* \cap \bigcap_{a \in \mathcal{B}(F)} O(a) \subseteq R. \quad (5.49)$$

Proof: We first show that

$$(\exists a \in \mathcal{A} P_s^\sigma(O(a)) = 1 \ \& \ F \in \text{Op}_\sigma) \Rightarrow \forall a \in \mathcal{B}(A) : P_s^\sigma(O(a)) = 1. \quad (5.50)$$

Let a, b be such that $P_s^\sigma(O(a)) = 1$ and $P_s^\sigma(O(b)) < 1$. Then

$$\begin{aligned} EU_E^\sigma(b) &= \sum_{v \in O(a)} P_s^\sigma(v) \cdot u(v, b) < \sum_{v \in O(a) \cap O(b)} P_s^\sigma(v) \cdot u(v, a) \\ &+ \sum_{v \in O(a) \setminus O(b)} P_s^\sigma(v) \cdot u(v, a) = EU_E^\sigma(a). \end{aligned}$$

With $K_s = \{v \in \Omega \mid P_s^\sigma(v) > 0\}$ it follows that $b \notin \mathcal{B}(K_s)$, and by (4.37) that $b \notin \mathcal{B}(A)$. Hence, $b \in \mathcal{B}(A)$ implies $P_s^\sigma(O(b)) = 1$.

Let $F^+ := \bigcap_{a \in \mathcal{B}(A)} O(a)$. We first show that $F^* \cap F^+ \subseteq R$ implies $F +> R$. Let $\hat{\sigma} \in \mathcal{S}$ be such that $F \in \text{Op}_{\hat{\sigma}}$. We have to show that $P_s^{\hat{\sigma}}(R) = 1$. By (5.50) $P_s^{\hat{\sigma}}(F^+) = P_s^{\hat{\sigma}}(\bigcap_{a \in \mathcal{B}(A)} O(a)) = 1$ and by (4.32) $P_s^{\hat{\sigma}}(F^*) = 1$; hence $P_s^{\hat{\sigma}}(F^+ \cap F^*) = 1$, and it follows that $P_s^{\hat{\sigma}}(R) = 1$.

Next, we show $F +> R$ implies $F^* \cap F^+ \subseteq R$. Suppose that $F^* \cap F^+ \not\subseteq R$. Let $w \in F^* \cap F^+ \setminus R$. From condition $\forall v \in \Omega \exists \hat{\sigma} \in \mathcal{S} P_s^{\hat{\sigma}}(v) = 1$ it follows that there is a support problem $\hat{\sigma}$ such that $P_s^{\hat{\sigma}}(w) = 1$. As $w \in F^+$, it follows by (4.37) that $F \in \text{Op}_{\hat{\sigma}}$. Due to $F +> R$, it follows that $P_s^{\hat{\sigma}}(R) = 1$, in contradiction to $w \notin R$. ■

F^* is the equivalent to the common ground updated with F . In the context of a support problem, we can interpret an answer F as a *recommendation* to choose one of the action in $\mathcal{B}(F)$. We may say that the recommendation is *felicitous* only if all recommended actions are optimal. Hence, F^+ represents the information that follows from the felicity of the speech act of recommendation which is associated to the answer. It should also be mentioned that $\mathcal{B}(F) = \mathcal{B}(F^*)$ by Definition 4.30; hence $\bigcap_{a \in \mathcal{B}(F)} O(a) = \bigcap_{a \in \mathcal{B}(F^*)} O(a)$

It is not uninteresting to see that the expert assumption on its own does not guarantee that an utterance has non-trivial implicatures. There are sets \mathcal{S} in which the conditions of Lemma 5.6 hold but in which answers only implicate what they entail:

Proposition 5.7 *Let \mathcal{S} be a set of support problems with joint decision problem $\langle(\Omega, P_H), \mathcal{A}, u\rangle$. Let (S, H) be the canonical solution. Assume that for all $X \subseteq \Omega$, $X \neq \emptyset : \exists \sigma \in \mathcal{S} K_s^\sigma = X$ and $\exists a \in \mathcal{A} O(a) = X$. Then, for all $\sigma \in \mathcal{S}$ with $F \in \text{Op}_\sigma$ it holds $\forall R \subseteq \Omega : \langle \mathcal{S}, S, H \rangle \models F +> R \Leftrightarrow F^* \subseteq R$.*

Proof: Condition $\forall X \neq \emptyset \exists a \in \mathcal{A} O(a) = X$ trivially entails that E is an expert for all $\sigma \in \mathcal{S}$. Condition $\forall X \neq \emptyset \exists \sigma \in \mathcal{S} K_s^\sigma = X$ entails the second condition of Lem. 5.6: $\forall v \in \Omega \exists \sigma \in \mathcal{S} P_s^\sigma(v) = 1$. Then, let $F \in \text{Op}_\sigma$ and let a^* be such that $O(a^*) = F^*$; as $\mathcal{B}(F) = \mathcal{B}(F^*)$, it follows that $\bigcap \{O(a) \mid a \in \mathcal{B}(F)\} = \bigcap \{O(a) \mid a \in \mathcal{B}(F^*)\} = O(a^*) = F^*$. Hence, by Lem. 5.6, $F +> R$ iff $F^* \subseteq R$. ■

This proposition also shows that the conditions of Lemma 5.6 are less restrictive than they might seem to be.

6 The Fundamental Lemma

In this section we prove Lemma 2.3. For convenience, let us again repeat the definition of signalling games from Definition 2.1. A signalling game is a structure $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ for which: (1) Ω and Θ are non-empty finite sets; (2) $P(\cdot)$ is a probability distribution over Ω ; (3) $p(\cdot|v)$ is a probability distribution over Θ for every $v \in \Omega$; (4) \mathcal{F} and \mathcal{A} are respectively the speaker's and hearer's action sets; and (5) $u : \Omega \times \Theta \times \mathcal{F} \times \mathcal{A} \rightarrow \mathbb{R}$ is a shared utility function which can be decomposed such that $u(v, \theta, F, a) = u(v, a) - c(F)$ for some strictly positive function $c : \mathcal{F} \rightarrow \mathbb{R}^+$.

We first introduce Bayesian perfect equilibria and then prove Lemma 2.3. As mentioned before, it can be more convenient to calculate the Bayesian perfect equilibria than the Nash equilibria of a signalling game.

Definition 6.1 (Perfect Bayesian Equilibrium) *A strategy pair (S, H) is a perfect Bayesian equilibrium of a signalling game $\langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ iff:*

1. *For all S' and all θ with $\mu(\theta) > 0$ it is $\mathcal{E}_S(S'|\theta) \leq \mathcal{E}_S(S|\theta)$,*
2. *For all H' and all F with $\mu(F) > 0$ it is $\mathcal{E}_H(H'|F) \leq \mathcal{E}_H(H|F)$.*

The equilibrium is strict if we can replace \leq by $<$. It is weak if it is not strict.

We show that the Bayesian perfect equilibria are the same as Nash equilibria in the sense of Definition 2.2. For this we show that a hearer strategy H is a best response to a speaker strategy S iff $\mathcal{E}_H(H|F)$ is maximal for each F with

non-negative probability. For the following calculations, it should be noted that the payoff function u does not depend on θ . Hence, we could arbitrarily choose a θ_0 and keep it as a fixed argument of u . With $\mu(F)$ defined as before, it is:

$$\begin{aligned}
\mathcal{E}(H|S) &= \sum_{v \in \Omega} P(v) \sum_{\theta \in \Theta} p(\theta|v) \sum_{F \in \mathcal{F}} S(F|\theta) \sum_{a \in \mathcal{A}} H(a|F) u(v, \theta, F, a) \\
&= \sum_F \mu(F) \sum_v \sum_a H(a|F) u(v, \theta_0, F, a) \frac{P(v) \sum_{\theta \in \Theta} p(\theta|v) S(F|\theta)}{\mu(F)} \\
&= \sum_F \mu(F) \sum_v \mu_H(v|F) \sum_a H(a|F) u(v, \theta, F, a) \\
&= \sum_{F \in \mathcal{F}} \mu(F) \mathcal{E}_H(H|F). \tag{6.51}
\end{aligned}$$

Hence, for fixed speaker strategy S , $\mathcal{E}(H|S)$ becomes maximal iff $\mathcal{E}_H(H|F)$ is maximal for all F with $\mu(F) > 0$, i.e. for all F for which the probability of being received by the hearer is greater zero. Similarly, it can be shown that $\mathcal{E}(S|H) = \sum_{\theta} \mu(\theta) \mathcal{E}_S(S|\theta)$. Hence, the Bayesian perfect equilibria in the sense of Definition 6.1 are identical to the Nash equilibria in the sense of Definition 2.2.

We now turn to the proof of Lemma 2.3. For this, we first reformulate the hearer's expected utility in terms of the conditional probability of the speaker's type being θ given answer F . This allows us to derive an estimate of the maximal expected utility. Hence, let us consider the expected utility $\mathcal{E}_H(H|F)$ of a hearer strategy H after receiving signal F . With (2.2) and (2.4), we find:

$$\begin{aligned}
\mathcal{E}_H(H|F) &= \sum_a H(a|F) \frac{\sum_v P(v) \sum_{\theta} p(\theta|v) S(F|\theta)}{\mu(F)} u(v, \theta, F, a) \\
&= \frac{1}{\mu(F)} \sum_{\theta} S(F|\theta) \sum_a H(a|F) \sum_v P(v) p(\theta|v) u(v, \theta, F, a) \\
&= \frac{1}{\mu(F)} \sum_{\theta} S(F|\theta) \mu(\theta) \sum_a H(a|F) (\mathcal{E}_S(a|\theta) - c(F)) \\
&= \sum_{\theta} \frac{S(F|\theta) \mu(\theta)}{\mu(F)} \sum_a H(a|F) \mathcal{E}_S(a|\theta) - c(F)
\end{aligned}$$

Let's write

$$\mu_{\Theta|\mathcal{F}}(\theta|F) = \frac{S(F|\theta) \mu(\theta)}{\mu(F)} = \frac{S(F|\theta) \sum_w P(w) p(\theta|w)}{\sum_{\theta} S(F|\theta) \sum_w P(w) p(\theta|w)}. \tag{6.52}$$

This is the hearer's probability of the speaker type being θ given F . In the following, we will also use the short form $\mu(\theta|F)$ for $\mu_{\Theta|\mathcal{F}}(\theta|F)$. With this

abbreviation, we can summarise the result as follows:

$$\mathcal{E}_H(H|F) = \sum_{\theta} \mu(\theta|F) \sum_a H(a|F) \mathcal{E}_S(a|\theta) - c(F) \quad (6.53)$$

Let $M_{\theta} := \max_a \mathcal{E}_S(a|\theta)$. This is the maximal expected utility given θ . An action is *optimal* given θ if its expected utility is maximal. Hence, $\mathcal{E}_S(a|\theta) = M_{\theta}$ iff a is an element of the set $\mathcal{B}(\theta)$ of all actions with maximal expected utility, which has been defined in (2.8) as:

$$\mathcal{B}(\theta) = \{a \in \mathcal{A} \mid \forall b \in \mathcal{A} \mathcal{E}_S(b|\theta) \leq \mathcal{E}_S(a|\theta)\}. \quad (6.54)$$

Hence, for fixed θ we find:

1. If $H(a|F) > 0 \Rightarrow a \in \mathcal{B}(\theta)$, then

$$\sum_a H(a|F) \mathcal{E}_S(a|\theta) = M_{\theta}. \quad (6.55)$$

2. If $\exists a \notin \mathcal{B}(\theta) H(a|F) > 0$, then

$$\sum_a H(a|F) \mathcal{E}_S(a|\theta) < M_{\theta}. \quad (6.56)$$

It follows then from (6.53) that a strategy H is guaranteed to be optimal if $\mu(\theta|F) > 0$ entails for all θ that $(H(a|F) > 0 \Rightarrow a \in \mathcal{B}(\theta))$, i.e. $H(\mathcal{B}(\theta)|\theta) = 1$. As mentioned before, we implicitly assume that in (6.52) the denominator $\mu(F)$ of μ is greater zero. Hence, if $\mu(\theta) > 0$, i.e. if θ is assigned to the speaker with a positive probability, then it follows that $\forall \theta (S(F|\theta) > 0 \Rightarrow H(\mathcal{B}(\theta)|F) = 1)$ entails that $\mathcal{E}_H(H|F)$ is maximal.

These considerations lead to the following criteria. Let Θ^* be the set of all types θ for which $\exists v P(v) p(\theta|v) > 0$, and let $F \in \mathcal{F}$. Let (S, H) be a strategy pair which satisfies the following condition:

$$\forall \theta \in \Theta^* (S(F|\theta) > 0 \Rightarrow H(\mathcal{B}(\theta)|F) = 1). \quad (6.57)$$

Then it follows that H is a best response to F , i.e. for all hearer strategies H' it holds that $\mathcal{E}_H(H'|F) \leq \mathcal{E}_H(H|F)$. Furthermore, if H' is such that

$$\exists \theta \in \Theta^* \exists a \notin \mathcal{B}(\theta) (S(F|\theta) > 0 \wedge H'(a|F) > 0), \quad (6.58)$$

then $\mathcal{E}_H(H'|F) < \mathcal{E}_H(H|F)$. Equations (6.57) and (6.58) entail Lemma 2.3.

7 Implicatures and Ambiguity

In this section, we address a problem that is shared by all accounts that assume that disambiguation is achieved by maximising expected utilities of interpretations. This method of disambiguation is the central principle for explaining pragmatic phenomena in Prashant Parikh's framework of games of partial information (2001). His standard example is the following sentence showing a scope ambiguity:

- (2) a) Every ten minutes a man gets mugged in New York. (*A*)
b) Every ten minutes some man or other gets mugged in New York. (*F*)
c) Every ten minutes a particular man gets mugged in New York. (*F'*)

The sentence *A* is ambiguous between the interpretation in which it is always the same person which gets mugged (*R'*), and the interpretation in which it is a random sequence of people who gets mugged (*R*). Speaker and hearer have to coordinate their strategies such that the hearer arrives at the interpretation that the speaker had in mind. With *F* being the unambiguous sentence with meaning *R*, *F'* the unambiguous sentence with meaning *R'*, and ρ, ρ' the probabilities of *R, R'* respectively, we arrive at the game tree shown in Figure 4. First nature

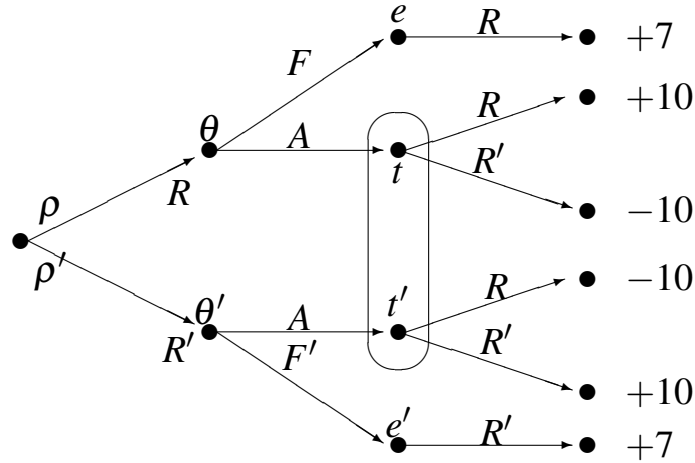


Fig. 4: Parikh's game tree for Example (2).

chooses between *R* and *R'*. If it has chosen *R*, then the speaker in situation θ has the choice between the unambiguous but more complex *F* and the ambiguous but simpler *A*. If nature has chosen *R'*, then the speaker in situation θ' has the choice between the unambiguous but more complex *F'* and again the ambiguous but simpler *A*. If the speaker chooses *F* or *F'*, then there is only one

interpretation which the hearer can choose. If the speaker chooses A , then the ambiguity of A leads to the choice between two different interpretations. The numbers at the end of the branches denote the shared utilities of speaker and hearer.

Prashant Parikh solves this game by calculating the Nash equilibria, and, if there is more than one Nash equilibrium, choosing the equilibrium which leads to the higher overall expected payoff, the so-called *Pareto* Nash equilibrium. It is easy to see that there are exactly two Nash equilibria in the situation of Figure 4: One Nash equilibrium (S, H) in which the speaker chooses F in θ and A in θ' , and in which the hearer interprets A by R' ; another Nash equilibrium (S', H') in which the speaker chooses F' in θ' and A in θ , and in which the hearer interprets A by R . As the probability ρ of it being always the same man who gets mugged is much lower than the probability ρ' , the first strategy will more often avoid the use of the complex formula F' , and hence lead to a higher overall expected utility. Hence, the first strategy pair (S, H) is the unique Pareto Nash equilibrium of this game, and the hearer will interpret A as meaning R . According to Parikh, this shows that the utterance of A *communicates with certainty* that R (Parikh, 1990), (Parikh, 2006)[p. 104].

Implicatures are explained by Parikh (2001) along the same lines. He assumes that an utterance is ambiguous between the literal meaning (A) and the literal meaning + implicature ($A + R$). The implicature $A +> R'$ is explained by the fact that for the Pareto Nash equilibrium (S, H) which solves the resulting game it holds that $H(A) = R'$. This account is principally different from the account provided in the Optimal Answer model. There, the solution (S, H) is calculated by backward induction,⁵ and the implicature is identified with the additional information that an utterance A provides about the speaker's information state, i.e. with $S^{-1}(A)$. But, although the two approaches differ here, the same predictions about disambiguation are made in the Optimal Answer model and in Parikh's model.

Our principal counterexample against the idea that ambiguities are resolved by choosing the more probable interpretation, and hence that this interpretation is thereby communicated with *certainty*, is the Doctor's Appointment example:

- (3) John is known to regularly consult two different doctors, physicians A and B . He consults A more often than B . S meets H and tells him:

S: John has a doctor's appointment at 4pm. He requests you to pick him up afterwards. (D)

⁵As we have shown in Lemma 4.5, the solution found by backward induction is always a Pareto Nash equilibrium.

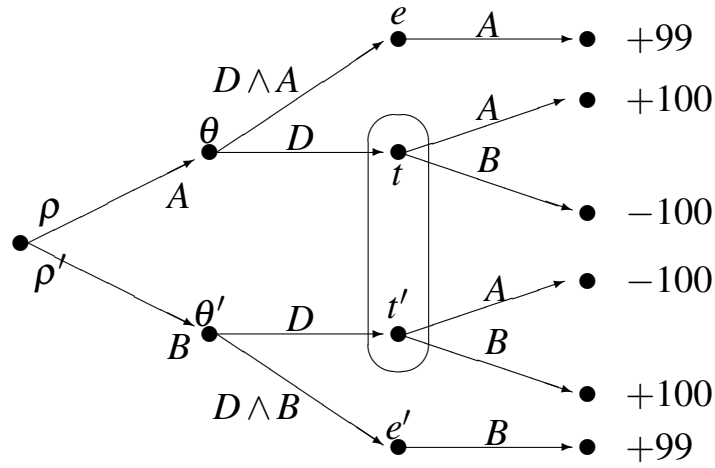


Fig. 5: The game tree for the Doctor's Appointment example.

Clearly, S fails to communicate that John waits at A's practice. Structurally, the situation is identical to the situation shown in Figur (4). In Figure 5, we see the game tree of the Doctor's Appointment example. The hearer has the choice between two different interpretations: A that John is waiting at A's practice, and B that John is waiting at B's practice. Hence, if rational interlocutors resolve the ambiguity by choosing the more probable interpretation in Figur (4), then in (3) they should resolve it in the same way. But in this case, D would have to communicate *with certainty* that John has an appointment with physician A, which is clearly not the case.

The problem does not only arise with Parikh's model. In the Doctor's Appointment example, backward induction predicts that D is an optimal assertion if the speaker knows that John is at A's practice but not if he knows that John is at B's practice. Hence, we are faced with the problem to explain why D is not an optimal assertion in the Doctor's Appointment example. This means, we have to explain why backward induction is ruled out as a principle of disambiguation in the Doctor's Appointment scenario. This problem leads us to the consideration of games with *noisy communication* and *efficient clarification requests*.

Let us consider the addressee's natural reaction in the Doctor's Appointment example (3). What would be the natural response to the directive (D) *John has a doctor's appointment at 4pm. He requests you to pick him up afterwards?* Most probably, the addressee would just ask where John is waiting, at A's or at B's practice. If the answering person S is cooperative and knows about John's whereabouts, then he will tell H where to pick up John. Hence, the natural response to the ambiguity is a clarification request c which will induce S to provide an answer which allows H to unambiguously choose an optimal action a afterwards. We call such clarification requests *efficient* if they come with low costs and force the speaker to provide an unambiguously optimal answer. We

add efficient clarification requests in the next section to our models.

8 Efficient clarification requests

Until now, our signalling games modelled situations in which the hearer has immediately to decide which action to choose after receiving an answer from the speaker. In this section, we add efficient clarification requests to the hearer's action set. In the context of noisy communication, the possibility to ask clarification requests has a considerable effect on the equilibria of a game.

Assume that the hearer follows a strategy H which does not involve any clarification requests. Assume further that the hearer knows the speaker strategy S and receives an answer A . We now consider the question: When is it reasonable for the hearer to change his strategy $H(\cdot | A)$ and ask a clarification request? For answering this question, we consider the set $Z(A)$ which consists of all pairs $\langle v, \theta \rangle$ which have positive probability, for which the speaker might answer A , and for which this might lead the hearer to choose a sub-optimal action. Clearly, for being optimal, S and H should be such that $Z(A)$ is empty. If it is not empty, the hearer can make it empty by changing H in such a way that he asks a clarification request whenever answer A occurs. We can consider this to be a *local* change in the sense that it only changes the hearer's strategy for answer A but leaves it unchanged for all other answers. Before we define local changes, we first introduce signalling games with *efficient clarification requests*:

Definition 8.1 Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a basic signalling game in the sense of Def. 2.1. Then \mathcal{G} is a signalling game with efficient clarification requests iff there exists an act $\mathbf{c} \in \mathcal{A}$ and a cost function $c : \mathcal{F} \cup \{\mathbf{c}\} \rightarrow \mathbb{R}^+$ which satisfy the following conditions:

1. (Efficiency): $u(v, \theta, F, \mathbf{c}) = \mu(\theta) M_\theta - (c(F) + c(\mathbf{c}))$,
with $M_\theta := \max_{a \in \mathcal{A} \setminus \{\mathbf{c}\}} \mathcal{E}_s(a | \theta)$ and $\mu(\theta)$ as in (2.3);
2. (Nominality): the cost function c is nominal;
3. (Avoid \mathbf{c}): $\forall A, B \in \mathcal{F} : |c(A) - c(B)| < c(\mathbf{c})$.

M_θ is the maximal expected utility given θ . Hence, *efficiency* means that \mathbf{c} achieves the maximal expected utility minus the costs of asking a clarification request. *Nominality* entails that these costs are positive but arbitrarily small in comparison to the utility of other actions. We will provide an exact definition of *nominality* in Section 11. (Avoid \mathbf{c}) says that if the speaker can find a more complex answer B which avoids a clarification request, then he will choose B rather than sticking to an answer A which he would have chosen otherwise.

As $\mathcal{B}(\theta) = \{a \in \mathcal{A} \mid \mathcal{E}_S(a|\theta) = M_\theta\}$, it follows that $\mathbf{c} \notin \mathcal{B}(\theta)$. With μ as in (6.52), it also follows that:

$$\mathcal{E}_H(\mathbf{c}|A) = \sum_{\theta} \mu(\theta|A) M_\theta - (c(A) + c(\mathbf{c})). \quad (8.59)$$

The proof is completely parallel to that of (6.53). As signalling games with efficient clarification requests are special cases of basic signalling games, it also follows that Lemma 2.3 remains valid.

Now we consider a situation in which speaker and hearer follow some strategy pair (S, H) which may violate condition (6.57). How can the hearer modify his strategy in order to achieve the best result? Assume that he receives answer A , then if there is a possibility that strategy H chooses a sub-optimal action, it is better for the hearer to ask a clarification request. This is the case if the following set $Z(A)$ is not empty:

$$Z(A) := \{\langle v, \theta \rangle \mid P(v) p(\theta|v) S(A|\theta) > 0 \wedge \exists a \notin \mathcal{B}(\theta) H(a|A) > 0\}. \quad (8.60)$$

$Z(A)$ is the set of all pairs of worlds v and speaker's types θ which have non-zero probability, for which the speaker answers A with non-zero probability, and for which H may choose a suboptimal action. It is convenient to collect the worlds v and the types θ that belong to $Z(A)$ in two separate sets:

$$Z^1(A) := \{v \in \Omega \mid \exists \theta \langle v, \theta \rangle \in Z(A)\}, \text{ and } Z^2(A) := \{\theta \in \Theta \mid \exists v \langle v, \theta \rangle \in Z(A)\}.$$

We show that if $Z(A) \neq \emptyset$, then the hearer strategy H is strictly dominated by a strategy $H_{\mathbf{c}}^A$ which is defined as follows:

$$H_{\mathbf{c}}^A(a|B) := \begin{cases} H(a|B) & \text{for } B \neq A \\ 1 & \text{for } a = \mathbf{c} \text{ and } B = A \end{cases}. \quad (8.61)$$

The strategy $H_{\mathbf{c}}^A$ is identical to H for all answers except A . For A it chooses a clarification request. We show the following proposition:

Proposition 8.2 *Let (S, H) be any strategy pair of a given signalling game \mathcal{G} with efficient clarification request \mathbf{c} . Assume that $\mu_{\mathcal{F}}(A) > 0$, see (2.3). With $\mu(v, \theta) := P(v) p(\theta|v)$, μ_H as in (2.2), and $\mu_{\Theta|\mathcal{F}}$ as in (6.52), the following equivalences hold:*

$$\begin{aligned} Z(A) \neq \emptyset &\Leftrightarrow \mu(Z(A)) > 0 \Leftrightarrow P(Z^1(A)) > 0 \Leftrightarrow \mu_H(Z^1(A)|A) > 0 \\ &\Leftrightarrow \mu_{\Theta|\mathcal{F}}(Z^2(A)|A) > 0. \end{aligned} \quad (8.62)$$

It holds:

1. *If $Z(A) \neq \emptyset$, then $H_{\mathbf{c}}^A$ strictly dominates H .*
2. *If $Z(A) = \emptyset$, then H strictly dominates $H_{\mathbf{c}}^A$.*

Proof: The equivalences follow by unfolding the definitions. By (6.53) it holds that $\mathcal{E}_H(H|A) = \sum_{\theta} \mu_{\Theta|\mathcal{F}}(\theta|A) \sum_a H(a|A) \mathcal{E}_s(a|\theta) - c(A)$. We can split the sum \sum_{θ} into $\sum_{\theta \in Z^2(A)}$ and $\sum_{\theta \notin Z^2(A)}$. Hence, if $Z(A) = \emptyset$, and therefore $\mu_{\Theta|\mathcal{F}}(Z^2(A)|A) = 0$, then

$$\begin{aligned} \mathcal{E}_H(H|A) &= \sum_{\theta \notin Z^2(A)} \mu_{\Theta|\mathcal{F}}(\theta|A) M_{\theta} - c(A) \\ &> \sum_{\theta \notin Z^2(A)} \mu_{\Theta|\mathcal{F}}(\theta|A) M_{\theta} - (c(A) + c(\mathbf{c})) = \mathcal{E}_H(H_{\mathbf{c}}^A|A). \end{aligned}$$

Hence, H strictly dominates $H_{\mathbf{c}}^A$. Next, assume that $Z(A) \neq \emptyset$. This entails that:

$$\sum_{\theta} \mu_{\Theta|\mathcal{F}}(\theta|A) \sum_a H(a|A) \mathcal{E}_s(a|\theta) < \sum_{\theta} \mu_{\Theta|\mathcal{F}}(\theta|A) M_{\theta}$$

Hence, it follows with (Nominality) that

$$\begin{aligned} \mathcal{E}_H(H|A) &= \sum_{\theta} \mu_{\Theta|\mathcal{F}}(\theta|A) \sum_a H(a|A) \mathcal{E}_s(a|\theta) - c(A) \\ &< \sum_{\theta} \mu_{\Theta|\mathcal{F}}(\theta|A) M_{\theta} - (c(A) + c(\mathbf{c})) = \mathcal{E}_H(H_{\mathbf{c}}^A|A). \end{aligned}$$

Hence, $H_{\mathbf{c}}^A$ strictly dominates H . ■

The proposition shows how the hearer can improve his strategy for answer A without changing his strategy for answers different from A . We will exploit this property in the next section. In the same section, we will also see that a clarification request is not always the best response in situations in which the old strategy H would lead to sub-optimal choices. Proposition 8.2 only says that reacting with a clarification request is better than sticking to a faulty strategy, but there may be other possibilities to improve the old strategy. For achieving this result, we have to have a closer look at $Z(A)$. But this is more interesting in the context of noisy speaker strategies. We introduce models with expected noise in the next section, and will take up our consideration of $Z(A)$ in Section 10.

9 Expected Noise

There exist quite a number of equilibrium refinements in game theory which try to spell out which equilibria are stable under the assumption that strategies are noisy. Among the most widely discussed equilibria which deal with noisy strategies are *trembling hand perfect* equilibria (Selten, 1975) and *proper* equilibria (Myerson, 1978). In the context of support problems, a trembling hand perfect equilibrium is a pair of mixed strategies (s, h) such that there exists a sequence $(s^k, h^k)_{k=0}^{\infty}$ of completely mixed strategies which converge to (s, h)

such that s is a best responses to each h^k and h to each s^k . A strategy is *completely mixed* if it chooses every possible action with positive probability. That (s, h) is robust against *small* mistakes is captured by the condition that s and h need only to be best responses if h^k and s^k come close to h and s . For proper equilibria it is assumed that the probability of mistakes depends on how good an action is. In our context, this means for a perturbed speaker strategy \tilde{S} that the speaker chooses an answer which is just second to an optimal answer with probability at most ε times the probability of an optimal answer, and an answer that is third to an optimal answers with a probability ε times the probability of an second best answer, etc. For both criteria, the probability and the kind of mistakes can be inferred from theory *internal* parameters, as e.g. from the set of available hearer actions, their expected utilities, and the speaker's set of signals. In linguistic pragmatics, in contrast to other applications of game theory, the phenomena are very close to the cognitive level. Hence a strong interaction between the behavioural level, represented by game theory, and the cognitive level is to be expected. We introduce expected noise models as a framework to introduce noise into the game theoretic models which is controlled by *external* causes. The representation of noise in expected noise models is therefore very little restricted. In other respects, we simplify the model by only considering perturbations of the speaker's strategy. This means, we always assume that the hearer finds his best response with certainty.

In order to motivate the following definition, assume that an interpreted signalling game σ is given. Assume further that the speaker follows strategy S . Then $S(\cdot | \sigma)$ will assign non-zero probability to certain forms $F \in \mathcal{F}$. They may form a proper sub-set \mathcal{O} of \mathcal{F} . We may call S^ε a noisy ε -approximation of S if $\sum_F |S(F | \sigma) - S^\varepsilon(F | \sigma)| = \varepsilon$. Then, for $S^\varepsilon(\cdot | \sigma)$ we can also collect all forms to which S^ε assign non-zero probability in a set \mathcal{N}^ε . The exact value of ε does not matter to us; hence, we abstract away from it and just keep the set of forms to which the S^ε assign non-zero probability. We assume that it is the same set for all ε . We call this set a *noise set*. Now, as we want to capture by these noise sets the perturbations resulting from cognitive sources, these sets may vary from support problem to support problem as the speaker's state varies from support problem to support problem. We therefore represent the perturbations by a function which maps \mathcal{S} to sets $\mathcal{N}_\sigma \subseteq \mathcal{F}$. This motivates the following definition:

Definition 9.1 (EN model) *Let \mathcal{S} be a set of interpreted support problems. Assume that the support problems $\langle \Omega, P_S, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ may only differ with respect to P_S . A model with expected noise, or EN model, is a triple $\langle \mathcal{S}, (\mathcal{O}_\sigma)_{\sigma \in \mathcal{S}}, (\mathcal{N}_\sigma)_{\sigma \in \mathcal{S}} \rangle$ for which*

1. $(\mathcal{O}_\sigma)_{\sigma \in \mathcal{S}}$ is a sequence of sets $\mathcal{O}_\sigma \subseteq \mathcal{F}$.

2. $(\mathcal{N}_\sigma)_{\sigma \in \mathcal{S}}$ is a sequence of sets $\mathcal{N}_\sigma \subseteq \mathcal{F}$.

In the following, we write $\langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ instead of $\langle \mathcal{S}, (\mathcal{O}_\sigma)_{\sigma \in \mathcal{S}}, (\mathcal{N}_\sigma)_{\sigma \in \mathcal{S}} \rangle$. In our applications, \mathcal{O}_σ is the set Op_σ of optimal answers of the canonical solution to the support problem σ .

If the hearer cannot distinguish between the elements of \mathcal{S} , then learning that the speaker produced a possibly noisy form F only provides him with the information that F is an element of the union of the \mathcal{N}_σ . Hence, we introduce the set:

$$\mathcal{N} := \bigcup_{\sigma} \mathcal{N}_\sigma. \quad (9.63)$$

It can be easily seen that the addition of efficient clarification requests in itself has no effect on the canonical solution. This changes when we consider *noisy communication*. We will see that the addition of noise and the availability of efficient clarification requests gives rise to a transformation (\tilde{S}, \tilde{H}) of the canonical solution (S, H) which is Pareto dominating all other strategies. In addition, the transformed solution is robust against the noise characterised by an EN model. A central role will be played by the sets $\tilde{\mathcal{B}}(A)$ and \mathcal{F}_{en} . $\tilde{\mathcal{B}}(A)$ is the set of all actions a which are optimal for all support problems σ for which A can occur as a noisy form:

$$\tilde{\mathcal{B}}(A) := \bigcap \{ \mathcal{B}(K_\sigma) \mid A \in \mathcal{N}_\sigma \} \text{ with } K_\sigma = \{v \mid P_s^\sigma(v) > 0\}. \quad (9.64)$$

\mathcal{F}_{en} then collects all $A \in \mathcal{F}$ for which $\tilde{\mathcal{B}}(A)$ is not empty:

$$\mathcal{F}_{\text{en}} := \{A \in \mathcal{N} \mid \tilde{\mathcal{B}}(A) \neq \emptyset\}. \quad (9.65)$$

We illustrate the meaning of these sets by a little example. Assume there are two support problems σ and σ' . Assume further that actions a and b are optimal in σ , and actions b and c are optimal in σ' . *Being optimal* has to be understood relative to the speaker's expectations P_s^σ ; hence, we mean by saying that a and b are optimal in σ that $EU_s^\sigma(a) = EU_s^\sigma(b) = \max\{EU_s^\sigma(a') \mid a' \in \mathcal{A}\}$, and therefore *being optimal* is equivalent to being an element of $\mathcal{B}(K_\sigma)$. Hence, it is $\mathcal{B}(K_\sigma) = \{a, b\}$ and $\mathcal{B}(K_{\sigma'}) = \{b, c\}$. Let us now assume that $S^\varepsilon(A|\sigma) > 0$ and $S^\varepsilon(A|\sigma') > 0$. What is the best response for the hearer? If he chooses a , then this may be sub-optimal as he cannot be sure that the actual support problem is really σ . The same problem arises with c . But if he chooses b , then he is save as b is an optimal action in both σ and σ' . Choosing b is also better than asking a clarification request as this comes with additional (nominal) costs. Now, in contrast, consider a situation in which actions a and b are optimal in σ , and actions c and d are optimal in σ' . If again $S^\varepsilon(A|\sigma) > 0$ and $S^\varepsilon(A|\sigma') > 0$, then it is now better to ask a clarification request as there is no best choice which

belongs to all $\mathcal{B}(K_\sigma)$. Hence, we see that if $\mathcal{F}_{\mathfrak{En}} \neq \emptyset$, then the hearer can safely choose an act in $\mathcal{F}_{\mathfrak{En}}$, otherwise he should react with a clarification request.

As mentioned before, \mathcal{O}_σ and \mathcal{N}_σ are representing speaker strategies S and their perturbed forms S^ε . The following definition makes explicit in which sense EN models represent these strategies:

Definition 9.2 Let $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ be an EN model with \mathcal{S} a set of support problems $\sigma = \langle \Omega, P_s, P_h, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$. For $X \subseteq \mathcal{F}$, we denote by Δ_X^* the set of all completely mixed strategies over X , i.e. Δ_X^* is the set of all probability distributions P over X for which $P(x) > 0$ iff $x \in X$. Then, we say that:

1. \mathfrak{En} represents a strategy S iff for all $\sigma \in \mathcal{S}$ $S(\cdot | \sigma) \in \Delta_{\mathcal{O}_\sigma}^*$;
2. \mathfrak{En} represents a noise strategy S^ε iff for all $\sigma \in \mathcal{S}$ $S^\varepsilon(\cdot | \sigma) \in \Delta_{\mathcal{N}_\sigma}^*$;
3. an arbitrary strategy \tilde{S} is an \mathfrak{En} strategy iff for all $\sigma \in \mathcal{S}$

$$\tilde{S}(\cdot | \sigma) \in \Delta_{\mathfrak{En}}^* := \Delta_{\mathcal{O}_\sigma}^* \cup \Delta_{\mathcal{N}_\sigma}^* \cup \Delta_{\mathcal{F}_{\mathfrak{En}}}^*. \quad (9.66)$$

Expected noise models are extensions of interpreted support problems. They represent the subjective level. Signalling games model the objective level, especially, objective success of communication is only definable for signalling games. Hence, in the following, we have again to consider the relation between signalling games and expected noise models. We repeat the definitions of the most important relations:

Definition 9.3 A signalling game \mathcal{G} supports an EN model $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ iff \mathcal{G} supports \mathcal{S} . By Definition 4.4, this means that $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ is such that $\Theta = \mathcal{S}$ and for all $\sigma = \langle \Omega, P_s, P_h, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$ it is $\mu_\Theta(\sigma) = \sum_v P(v) p(\sigma | v) > 0$. We say that \mathcal{G} fully supports \mathfrak{En} iff all P_s^σ are fully reliable; it reliably supports \mathfrak{En} iff all P_s^σ are reliable; and it weakly supports \mathfrak{En} iff all P_s^σ are truth preserving.

9.1 The canonical solution

We consider situations in which the speaker may follow some perturbed strategy S^ε , and the hearer a strategy H . There may exist a support problem σ for which the speaker choose an answer A with probability greater zero to which the hearer may respond with a sub-optimal action if he follows H . Assume that this perturbed strategy S^ε is known to the hearer, and that he is in a situation in which he receives answer A . How does the hearer have to change his strategy $H(\cdot | A)$ in order to achieve maximal expected payoff? We introduce an operation which changes the hearer's response to A but leaves it unchanged for all

other answers:

$$H_X^A(a|B) := \begin{cases} H(a|B) & \text{if } B \neq A, \\ |X|^{-1} & \text{if } B = A \wedge a \in X. \end{cases} \quad (9.67)$$

This operation turns strategy H into a strategy H_X^A which is identical to H for all answers except A , and for A it chooses each of the elements of X with equal probability. The strategy H_c^A defined in (8.61) is the special case in which the old response to A is replaced by the clarification request c . We also consider the case in which the old response to A is replaced by $\tilde{\mathcal{B}}(A)$. We write:

$$H_c^A := H_{\{c\}}^A \quad \text{and} \quad H^A := H_{\tilde{\mathcal{B}}(A)}^A. \quad (9.68)$$

It is quite intuitive that the hearer can optimise his strategy by changing $H(\cdot|A)$ to H_c^A if $\tilde{\mathcal{B}}(A) = \emptyset$, and by changing it to H^A if $\tilde{\mathcal{B}}(A) \neq \emptyset$. If (S, H) is the canonical solution to \mathcal{S} , then, by applying these operations systematically, we arrive at a new *canonical solution* to the expected noise model. Its definition is provided in (9.69) and (9.71).

Definition 9.4 (Canonical Solution) *Let \mathcal{S} be a set of support problems with canonical solution (S, H) . Let $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ be an expected noise model which represents S . Then, we define the canonical extension (\bar{S}, \bar{H}) to \mathfrak{En} as follows:*

$$\bar{H}(\cdot|A) = \begin{cases} H^A, & \text{if } A \in \mathcal{F}_{\mathfrak{En}}, \\ H_c^A & \text{if } A \notin \mathcal{F}_{\mathfrak{En}}. \end{cases} \quad (9.69)$$

For the speaker let $\bar{c}_\sigma := \min\{c(A) \mid A \in \mathcal{N}_\sigma \cap \mathcal{F}_{\mathfrak{En}}\}$, and:

$$\text{Op}_\sigma^{\mathfrak{En}} := \{A \in \mathcal{N}_\sigma \cap \mathcal{F}_{\mathfrak{En}} \mid c(A) = \bar{c}_\sigma\}. \quad (9.70)$$

Then, \bar{S} is defined by:

$$\bar{S}(A|\sigma) = \begin{cases} |\text{Op}_\sigma^{\mathfrak{En}}|^{-1} & \text{if } A \in \text{Op}_\sigma^{\mathfrak{En}} \\ 0 & \text{otherwise} \end{cases}. \quad (9.71)$$

We can show an equivalent to Lemma 2.3:

Lemma 9.5 *Let \mathcal{S} be a set of support problems with canonical solution (S, H) . Let \mathcal{G} be a signalling game which supports \mathcal{S} . Let, furthermore, \mathfrak{En} be an expected noise model which represents S . Then, the canonical solution (\bar{S}, \bar{H}) always exists, and it is a Bayesian perfect equilibrium of \mathcal{G} . In addition, if we treat nominal costs as zero, then (\bar{S}, \bar{H}) is Pareto dominating all other strategy pairs.*

This is the best result we can hope to achieve. We cannot exclude the possibility that there are signalling strategies which are more efficient than (\bar{S}, \bar{H}) .

For example, we may conceive an artificial signalling strategy for which the speaker says ‘A’ and snips with his fingers whenever he wants to say that there is a garage round the corner, and behaves exactly as if following \bar{S} in all other situations. Then, this strategy is arguably more cost efficient than \bar{S} . As our framework does not exclude such artificial signalling strategies, we cannot in general prove that (\bar{S}, \bar{H}) is Pareto dominating all other strategies.

9.2 Implikatures in EN models

We now consider the implicatures of an *ENmodel* $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$. As \mathcal{O}_σ and \mathcal{N}_σ are only there in order to represent strategies and their perturbations, they do not change the set of signalling games which support \mathcal{S} . As implicatures are an objective notion in our framework, and the objective level is described by signalling games, it follows that the implicatures of a signal F relative to a strategy S and an EN model \mathfrak{En} are the same as its implicatures relative to S and \mathcal{S} . Hence, we set for $F \in \mathcal{F}$ for which $\exists \sigma \in \mathcal{S} S(F|\sigma) > 0$, and $R \subseteq \Omega$:

$$\langle \mathfrak{En}, S, H \rangle \models F +> R \iff \langle \mathcal{S}, S, H \rangle \models F +> R. \quad (9.72)$$

By Definition 5.4, this is equivalent to:

$$\langle \mathfrak{En}, S, H \rangle \models F +> R \iff \forall \mathcal{G} \in \text{Supp}(\mathcal{S}) \langle \mathcal{G}, S, H \rangle \models F +> R. \quad (9.73)$$

Here, $\text{Supp}(\mathcal{S})$ is the set of all signalling games \mathcal{G} which support \mathcal{S} .

Let $\mathfrak{En}_0, \mathfrak{En}_1$ be two EN models which represent the same strategy pair (S, H) . Hence, $\mathcal{S}^{\mathfrak{En}_0} = \mathcal{S}^{\mathfrak{En}_1}$, and for each $\sigma \in \mathcal{S}^{\mathfrak{En}_i}$ it holds that $\mathcal{O}_\sigma^{\mathfrak{En}_i} = \{F \mid S(F|\sigma) > 0\}$. Then, Lemma 5.3 implies that:

$$\langle \mathfrak{En}_0, S, H \rangle \models F +> R \iff \langle \mathfrak{En}_1, S, H \rangle \models F +> R. \quad (9.74)$$

If (\bar{S}, \bar{H}) is the *canonical solution* to \mathfrak{En} , we arrive with (5.43) at:

$$\langle \mathfrak{En}, \bar{S}, \bar{H} \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (F \in \text{Op}_\sigma^{\mathfrak{En}} \Rightarrow P_s^\sigma(R) = 1). \quad (9.75)$$

Finally, we note a consequence of the definition for perturbed strategies S^ε . Let $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ be an EN model which represents S and S^ε . Let H be an arbitrary hearer strategy. If F is such that $\exists \sigma S^\varepsilon(F|\sigma) > 0$ and $R \subseteq \Omega$, then we find again with (5.43) that:

$$\langle \mathfrak{En}, S, H \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (F \in \mathcal{O}_\sigma \Rightarrow P_s^\sigma(R) = 1), \quad (9.76)$$

and

$$\langle \mathfrak{En}, S^\varepsilon, H \rangle \models F +> R \iff \forall \sigma \in \mathcal{S} (F \in \mathcal{N}_\sigma \Rightarrow P_s^\sigma(R) = 1). \quad (9.77)$$

10 On the equilibrium properties of the canonical solution

Lemma 9.5 states that the best strategy pair that speaker and hearer can adopt in EN models is the *canonical* strategy pair defined in Section 9. More precisely, it states that the canonical strategy is a Bayesian perfect equilibrium, and, if we ignore nominal costs, it is even Pareto dominating all other strategies. The hearer's part of the canonical strategy is defined in (9.67), which in turn can be defined in terms of H_c^A . The goal of this section is to prove Lemma 9.5. In addition, it contains some more fine grained characterisations of the canonical strategy.

Let $\mathcal{G} = \langle \Omega, \Theta, P, p, \mathcal{F}, \mathcal{A}, u \rangle$ be a signalling game which represents \mathcal{S} , and let $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ be an EN model which represents a strategy S and the ε -approximations S^ε . Following our procedure in Section 8, we consider the following set:

$$Z_H^\varepsilon(A) := \{ \langle v, \sigma \rangle \mid P(v) p(\sigma|v) S^\varepsilon(A|\sigma) > 0 \wedge \exists a \notin \mathcal{B}(\sigma) H(a|A) > 0 \}. \quad (10.78)$$

For $\varepsilon = 0$, this corresponds to the set $Z(A)$ defined in (8.60). $Z_H^\varepsilon(A)$ is the set of all pairs of worlds v and support problems σ which have non-zero probability, for which the speaker answers A with non-zero probability, and for which H may choose a suboptimal action. In Proposition 8.2, we have shown that the hearer can improve if he reacts with a clarification request. Now we see that he can improve even more if he distinguishes between answers A which are elements of \mathcal{F}_{en} and answers A which are not elements of \mathcal{F}_{en} . We first show how $\tilde{\mathcal{B}}(A)$ and $Z_H^\varepsilon(A)$ are related to each other:

Proposition 10.1 *Let \mathcal{G} be a signalling game which fully supports the EN model $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$. Assume that \mathfrak{En} represents S and the ε -approximations S^ε . Let $A \in \mathcal{N}$. Then:*

$$\tilde{\mathcal{B}}(A) = \emptyset \Leftrightarrow \forall H (H(\mathbf{c}|A) < 1 \Rightarrow Z_H^\varepsilon(A) \neq \emptyset). \quad (10.79)$$

Proof: We first prove “ \Rightarrow ”: Assume that $\tilde{\mathcal{B}}(A) = \emptyset$. Then, by definition:

$$\forall a \in \mathcal{A} \setminus \{\mathbf{c}\} \exists \sigma, \sigma' : A \in \mathcal{N}_\sigma \cap \mathcal{N}(\sigma') \wedge a \notin \mathcal{B}(K_\sigma) \cap \mathcal{B}(K_{\sigma'}).$$

As P_s^σ is fully reliable for each $\sigma \in \mathcal{S}$, it follows that $\mathcal{B}(K_\sigma) = \mathcal{B}(\sigma)$ and $\mathcal{B}(K_{\sigma'}) = \mathcal{B}(\sigma')$. Let $a \in \mathcal{A} \setminus \{\mathbf{c}\}$, and let H be any hearer strategy with $H(a|A) > 0$. Then, there exists σ such that $A \in \mathcal{N}_\sigma$ and $a \notin \mathcal{B}(\sigma)$. As \mathfrak{En} is supported by \mathcal{G} and S^ε represented by \mathfrak{En} , it follows that $P(v) p(\sigma|v) S^\varepsilon(A|\sigma) > 0$; hence we can find a v such that $\langle v, \sigma \rangle \in Z_H^\varepsilon(A)$ and an a with $H(a|A) > 0 \wedge a \notin \mathcal{B}(\sigma)$. Therefore $Z_H^\varepsilon(A) \neq \emptyset$.

“ \Leftarrow ”: Assume that $\tilde{\mathcal{B}}(A) \neq \emptyset$. Let $a \in \tilde{\mathcal{B}}(A)$ and set $H(a|A) = 1$. Assume that $P(v) p(\sigma|v) S^\varepsilon(A|\sigma) > 0$. As P_s^σ is fully reliable for each $\sigma \in \mathcal{S}$, it follows

by definition of $\tilde{\mathcal{B}}(A)$ that $a \in \mathcal{B}(\sigma)$. Hence, $\langle v, \sigma \rangle \notin Z_H^\varepsilon(A)$. As v and σ are arbitrary, it follows that $Z_H^\varepsilon(A) = \emptyset$. ■

This shows how we can improve over Proposition 8.2: Only if $\mathcal{B}(A) = \emptyset$ the hearer reacts with a clarification request, else he chooses an act from $\mathcal{B}(A)$. $\mathcal{B}(A) = \emptyset$ is equivalent to $A \in \mathcal{F}_{\text{en}}$. The next proposition tells us how expected utilities behave if the hearer changes his strategy from $H(\cdot|A)$ to H_X^A .

Proposition 10.2 *Let \mathcal{G} be a signalling game which fully supports the EN model $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$. Assume that \mathfrak{En} represents S and the ε -approximations S^ε . Let $A \in \mathcal{N}$. Let \mathcal{E}_H be the hearer's expected utility if the speaker follows strategy S , $\mathcal{E}_H^\varepsilon$ be the hearer's expected utility if the speaker follows strategy S^ε , and \mathcal{E}_H^\sim be the hearer's expected utility if the speaker follows some other strategy \tilde{S} . All the expected utilities are defined relative to the probabilities in \mathcal{G} . Let $M_\sigma = \max_{a \in \mathcal{A} \setminus \{c\}} \mathcal{E}_s(a|\sigma)$, and let μ be as in (6.52). In the following equations, let $X \subseteq \mathcal{A}$ be such that $c \notin X$. Then:*

1. $\mathcal{E}_H(H|A) = \mathcal{E}_H(H_X^A|A) = \sum_\sigma \mu(\sigma|A) M_\sigma - c(A)$ for $X \subseteq \tilde{\mathcal{B}}(A)$,
2. $\mathcal{E}_H^\varepsilon(H_X^A|A) < \mathcal{E}_H^\varepsilon(H^A|A) = \sum_\sigma \mu^\varepsilon(\sigma|A) M_\sigma - c(A)$ for $X \setminus \tilde{\mathcal{B}}(A) \neq \emptyset$,
3. Let \tilde{S} be given with $\tilde{S}(\mathcal{F}_{\text{en}}|\sigma) = 1$ for all σ , and let $X \setminus \tilde{\mathcal{B}}(A) \neq \emptyset$, then

$$\mathcal{E}_H^\sim(H_X^A|A) < \mathcal{E}_H^\sim(H^A|A) = \sum_\sigma \mu(\sigma|A) M_\sigma - c(A).$$

Proof: We first show that $\mathcal{E}_H(H|A) = \mathcal{E}_H(H_X^A|A)$ for $X \subseteq \tilde{\mathcal{B}}(A)$: By (6.53) $\mathcal{E}_H(H|A) = \sum_\sigma \mu(\sigma|A) \sum_a H(a|A) \mathcal{E}_s(a|\sigma) - c(A)$; from (6.57) it follows that $H(a|A) > 0 \Rightarrow \mathcal{E}_s(a|\sigma) = M_\sigma$; hence $\mathcal{E}_H(H|A) = \sum_\sigma \mu(\sigma|A) M_\sigma - c(A)$; as $X \subseteq \tilde{\mathcal{B}}(A)$, it follows again with (6.53) and (6.57) that

$$\mathcal{E}_H(H_X^A|A) = \sum_\sigma \mu(\sigma|A) \sum_{a \in X} H_X^A(a|A) \mathcal{E}_s(a|\sigma) - c(A) = \sum_\sigma \mu(\sigma|A) M_\sigma - c(A).$$

Next, we turn to $\mathcal{E}_H^\varepsilon(H_X^A|A) < \mathcal{E}_H^\varepsilon(H^A|A)$ for all X with $\tilde{\mathcal{B}}(A) \subsetneq X$: By (6.53), $\mathcal{E}_H^\varepsilon(H_X^A|A) = \sum_\sigma \mu^\varepsilon(\sigma|A) \sum_a H_X^A(a|A) \mathcal{E}_s(a|\sigma) - c(A)$; we can divide \sum_σ into the sum over the set $M_0 = \{\sigma \mid H_X^A(\mathcal{B}(\sigma)|A) = 1\}$ plus the sum over the set $M_1 = \{\sigma \mid H_X^A(\mathcal{B}(\sigma)|A) < 1\}$; as $\tilde{\mathcal{B}}(A) \subsetneq X$, the second set is not empty; as $H^A(a|A) > 0 \Rightarrow a \in \tilde{\mathcal{B}}(A)$, it follows for M_1 that

$$\begin{aligned} \sum_{M_1} \mu^\varepsilon(\sigma|A) \sum_a H_X^A(a|A) \mathcal{E}_s(a|\sigma) - c(A) &< \sum_{M_1} \mu^\varepsilon(\sigma|A) M_\sigma - c(A) = \\ &= \sum_{M_1} \mu^\varepsilon(\sigma|A) \sum_a H^A(a|A) \mathcal{E}_s(a|\sigma) - c(A). \end{aligned}$$

By (6.57) it follows that equality holds if we replace M_1 by M_0 . This proves the claim.

Finally, the proof of $\mathcal{E}_H^\sim(H_X^A|A) < \mathcal{E}_H^\sim(H^A|A) = \sum_\sigma \mu(\sigma|A) M_\sigma - c(A)$ is almost identical to the previous case. ■

In order to improve the readability of formulas, we use the abbreviation $NC(s)$ for expressing the fact that a speaker strategy s is not optimal but differs from an optimal strategy s' only by a positive term with nominal costs; in addition, these nominal costs can only be reduced by a change of the speaker strategy s , not by a change of the hearer strategy. That is, if we write $EU(s'|h) = EU(s|h) - NC(s)$, we mean that $EU(s'|h) < EU(s|h)$ such that first $EU(s'|h) - EU(s|h)$ is nominal, and second there is no strategy h' for which $EU(s|h') > EU(s|h)$. We find:

Proposition 10.3 *Let \mathcal{S} be a set of support problems with canonical solution (S, H) . Let \mathcal{G} be a signalling game which supports \mathcal{S} . Let, furthermore, $\mathfrak{En} = \langle \mathcal{S}, \mathcal{O}_\sigma, \mathcal{N}_\sigma \rangle$ be an expected noise model which represents S , and let (\bar{S}, \bar{H}) be its canonical solution. Then:*

1. $EU(\bar{S}|\bar{H}) = \sum_\sigma \mu_\sigma(\sigma) (M_\sigma - \sum_A \bar{S}(A|\sigma) c(A));$
2. $EU(S^\varepsilon|\bar{H}) = EU(\bar{S}|\bar{H}) - NC(S^\varepsilon).$

If for all $\sigma \in \mathcal{S}$ $\mathcal{O}_\sigma \subseteq \mathcal{F}_{\text{en}}$, then

3. $EU(S|H) = EU(S, \bar{H}) = EU(\bar{S}|\bar{H}) - NC(S).$

Furthermore, if \tilde{S} is such that $\exists \sigma \tilde{S}(\mathcal{F}_{\text{en}}|\sigma) < 1$, then

4. $EU(\tilde{S}, \bar{H}) = EU(\bar{S}|\bar{H}) - NC(\tilde{S})$

Proof: 1) By definition, $\bar{S}(A|\sigma) > 0$ implies $\bar{H}(a|A) > 0 \Rightarrow a \in \mathcal{B}(\sigma)$. Then, (6.55) and (2.1) imply that $EU(\bar{S}|\bar{H}) = \sum_\nu P(\nu) \sum_\sigma p(\sigma|\nu) \sum_A \bar{S}(A|\sigma) (M_\sigma - c(A))$, which equals $\sum_\sigma \mu_\sigma(\sigma) (M_\sigma - \sum_A \bar{S}(A|\sigma) c(A))$.

2) Let $\mu^\varepsilon(A) := \sum_\nu P(\nu) \sum_\sigma p(\sigma|\nu) S^\varepsilon(A|\sigma)$. If $\mu^\varepsilon(A|\sigma) > 0$ and $\tilde{\mathcal{B}}(A) = \emptyset$, then $\bar{H}(\mathbf{c}|A) = 1$, i.e. the hearer will react to A with a clarification request \mathbf{c} . If $\mu^\varepsilon(A|\sigma) > 0$ and $\tilde{\mathcal{B}}(A) \neq \emptyset$, then S^ε will produce higher costs than \bar{S} , iff $c(A)$ is more costly than \bar{c}_σ . Hence, with $\mu_\sigma(\sigma) = \sum_\nu P(\nu) p(\sigma|\nu)$, we arrive at:

$$\begin{aligned} EU(S^\varepsilon|\bar{H}) &= EU(\bar{S}|\bar{H}) - \left(c(\mathbf{c}) \mu^\varepsilon(\{A \mid \tilde{\mathcal{B}}(A) = \emptyset\}) + \right. \\ &\quad \left. + \sum_\sigma \mu_\sigma(\sigma) \sum_A S^\varepsilon(A|\sigma) (c(A) - \bar{c}_\sigma) \right) \\ &= EU(\bar{S}|\bar{H}) - NC(S^\varepsilon). \end{aligned} \tag{10.80}$$

This proves the second claim.

3) The first equation is trivially true. The second equation follows similarly to 2) as $S(A|\sigma) > 0$ implies $c(A) \geq \bar{c}_\sigma$; hence:

$$EU(S|\bar{H}) = EU(\bar{S}|\bar{H}) - \sum_\sigma \mu(\sigma) \sum_A S(A|\sigma) (c(A) - \bar{c}_\sigma). \tag{10.81}$$

This again proves the claim.

4) Assume that $\exists \sigma \tilde{S}(\mathcal{F}_{\text{en}}|\sigma) < 1$. We split \mathcal{S} into $M_0 = \{\sigma \mid \tilde{S}(\mathcal{F}_{\text{en}}|\sigma) = 1\}$ and $M_1 = \{\sigma \mid \tilde{S}(\mathcal{F}_{\text{en}}|\sigma) < 1\}$. Then, for each $\sigma \in M_1$, we split \mathcal{F} into $F_0^\sigma = \{A \in \mathcal{F}_{\text{en}} \mid \tilde{S}(A|\sigma) > 0\}$ and $F_1^\sigma = \{A \in \mathcal{F} \setminus \mathcal{F}_{\text{en}} \mid \tilde{S}(A|\sigma) > 0\}$. As $\exists \sigma \tilde{S}(\mathcal{F}_{\text{en}}|\sigma) < 1$, it follows that $\exists \sigma \tilde{S}(F_0^\sigma|\sigma) > 0$. Then, let \tilde{S}' be the strategy which results from replacing each $A \in F_0^\sigma$ by a $B \in \mathcal{N}_\sigma \cap \mathcal{F}_{\text{en}}$. Then, clearly, $EU(\tilde{S}|\bar{H}) < EU(\tilde{S}'|\bar{H})$, and by definition $EU(\tilde{S}'|\bar{H}) = EU(\tilde{S}|\bar{H}) - NC(\tilde{S}')$. As the difference between $EU(\tilde{S}|\bar{H})$ and $EU(\tilde{S}'|\bar{H})$ is only nominal, it also follows that $EU(\tilde{S}|\bar{H}) = EU(\tilde{S}|\bar{H}) - NC(\tilde{S})$. ■

With these preparations, we can finally show:

Proof of Lemma 9.5: From the third claim of Prop. 10.2, it follows that the hearer has no better strategy against \bar{S} than \bar{H} , in particular, \bar{H} satisfies the Bayesian condition. From the fourth claim of Prop. 10.3 it follows that the speaker prefers strategies \tilde{S} with $\tilde{S}(\mathcal{F}_{\text{en}}|\sigma) = 1$ over strategies \tilde{S} with $\tilde{S}(\mathcal{F}_{\text{en}}|\sigma) < 1$. By definition, strategies \tilde{S} which satisfy $\tilde{S}(\mathcal{F}_{\text{en}}|\sigma) = 1$ cannot be better than \bar{S} against \bar{H} . Hence, it follows that (\bar{S}, \bar{H}) is a Bayesian perfect equilibrium of \mathcal{G} . From the first claim of Prop. 10.3, it immediately follows that (\bar{S}, \bar{H}) is Pareto dominating all other strategy pairs if nominal costs are treated as zero. ■

11 Nominality

In this section we supplement a precise definition of *nominal* costs. As the technical details are of minor interest to the purposes of the present paper, we present only the bare essentials.

Definition 11.1 (Nominality) Let u be a function which takes arguments $\mathbf{a} = \langle a_1, \dots, a_n \rangle$. Let u_1 and u_2 be two functions for which $u(\mathbf{a}) = u_1(\mathbf{a}) + u_2(\mathbf{a})$. By saying that u_2 is nominal with respect to u , we say that for any continuous function f and arguments \mathbf{a}, \mathbf{b} the inequality $f(u(\mathbf{a})) \leq f(u(\mathbf{b}))$ means that

$$\lim_{k \rightarrow 0^+} \text{sgn}(f(u_1(\mathbf{b}) + k u_2(\mathbf{b})) - f(u_1(\mathbf{a}) + k u_2(\mathbf{a}))) \geq 0. \quad (11.82)$$

In this formula, $k \rightarrow 0^+$ means that we only consider sequences of $k > 0$ which converge to 0. The signum function sgn is defined as follows:

$$\text{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

One should keep in mind that there are continuous f for which the limit in (11.82) is not defined. But it is always defined for constant, linear, or monotonic continuous functions f . For linear functions, (11.82) is equivalent to:

$$f(u_1(\mathbf{a})) < f(u_1(\mathbf{b})) \vee (f(u_1(\mathbf{a})) = f(u_1(\mathbf{b})) \wedge f(u_2(\mathbf{a})) < f(u_2(\mathbf{b}))). \quad (11.83)$$

As example, we consider the speaker's expected utility of uttering F given a hearer strategy H for a support problem $\langle \Omega, P_s, P_H, \mathcal{F}, \mathcal{A}, u, c, \llbracket \cdot \rrbracket \rangle$. The utility function u is of the form $u(v, F, a) = u(v, a) + c(F)$ with nominal c . Hence, we set $u_1(a, F, a) = u(v, a)$ and $u_2(a, F, a) = -c(F)$. The speaker's expected utility is defined as:

$$EU_s(F) = \sum_{v \in \Omega} P_s(v) \sum_{a \in \mathcal{B}(F)} H(a|F) u(v, F, a). \quad (11.84)$$

As EU_s is linear, we find:

$$EU_s^k(F) := EU_s^0(F) - k c(F), \quad (11.85)$$

Hence, the nominality of c entails that for all $F_0, F_1 \in \mathcal{F}$:

$$EU_s(F_0) \leq EU_s(F_1) :\Leftrightarrow \lim_{k \rightarrow 0^+} \text{sgn} \left(EU_s^k(F_1) - EU_s^k(F_0) \right) \geq 0. \quad (11.86)$$

Clearly, it follows that

$$EU_s^0(F_0) < EU_s^0(F_1) \Rightarrow EU_s(F_0) < EU_s(F_1). \quad (11.87)$$

For clarification requests we have to generalise the definition slightly. Expected utilities with clarification requests divide into a non-nominal term which depends on the speakers signal A , and a nominal term which is the sum of the costs for uttering A and the costs due to the clarification request \mathbf{c} . Hence, we generalise (11.86) as follows: For finite sets $X \subseteq \text{dom } \mathbf{c}$ we set

$$c(X) := \sum_{F \in X} c(F), \text{ and } EU_s^k(F_i, X) = EU_s^k(F_i) - k c(X). \quad (11.88)$$

Then, $EU_s(F_0, X_0) \leq EU_s(F_1, X_1)$ iff:

$$\lim_{k \rightarrow 0^+} \text{sgn} \left(EU_s^k(F_1, X_1) - EU_s^k(F_0, X_0) \right) \geq 0. \quad (11.89)$$

For example, F_0 and F_1 may be two possible utterances, and $X_0 = \{\mathbf{c}\}$ and $X_1 = \emptyset$. It follows by definition that

$$EU_s^0(F_0) < EU_s^0(F_1) \Rightarrow EU_s(F_0, X_0) < EU_s(F_1, X_1). \quad (11.90)$$

References

- Benz, A. (2006). Utility and Relevance of Answers. In Benz, A., Jäger, G., and van Rooij, R., editors, *Game Theory and Pragmatics*, pages 195–214. Palgrave Macmillan, Basingstoke.
- Benz, A. (2007). On Relevance Scale Approaches. In Puig-Waldmüller, E., editor, *Proceedings of the Sinn und Bedeutung 11*, pages 91–105.

- Benz, A. (2008). How to Set Up Normal Optimal Answer Models. Ms, ZAS, Berlin.
- Benz, A. and van Rooij, R. (2007). Optimal assertions and what they implicate: a uniform game theoretic approach. *Topoi - an International Review of Philosophy*, 27(1):63–78.
- Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. PhD thesis, Universiteit van Amsterdam.
- Grice, H. P. (1957). Meaning. *Philosophical Review*, 66:377–388.
- Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge MA.
- Groenendijk, J. and Stockhof, M. (1991). Dynamic predicate logic. *Linguistics & Philosophy*, 14:39–100.
- Jäger, G. and Ebert, C. (2009). Pragmatic Rationalizability. In Riester, A. and Solstad, T., editors, *Proceedings of Sinn und Bedeutung*, volume 13.
- Lewis, D. (2002). *Convention*. Blackwell Publishers, Oxford. First published by Harvard University Press 1969.
- Myerson, R. (1978). Refinements of the Nash equilibrium concept. *International Journal of game theory*, 7:73–80.
- Parikh, P. (1990). Situations, Games, and Ambiguity. In Cooper, R., Mukai, K., and Perry, J., editors, *Situation Theory and its Applications*, volume 1, pages 449–469. CSLI Lecture Notes, Stanford, CA.
- Parikh, P. (2001). *The Use of Language*. CSLI Publications, Stanford.
- Parikh, P. (2006). Pragmatics and Games of Partial Information. In Benz, A., Jäger, G., and van Rooij, R., editors, *Game Theory and Pragmatics*, pages 101–121. Palgrave Macmillan, Basingstoke.
- Pearle, J. (2000). *Causality - Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.
- Savage, L. (1972). *The Foundations of Statistics*. Dover Publications, New York. revised and enlarged version; first published by John Wiley & Sons, 1954.
- Selten, R. (1975). Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory*, 4:25–55.

ZAS Papers in Linguistics were originally published by the Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Typologie und Universalienforschung (FAS, Research Center for General Linguistics, Typology and Universals). The Center is now known as *Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung* (ZAS) under the auspices of the Deutsche Forschungsgemeinschaft (The German Research Foundation) and the State of Berlin. The Center currently has research projects in syntax, semantics, morphology, phonology, phonetics as well as language contact and language acquisition. ZAS provides a forum for the exchange of ideas in the academic community of the Berlin area through lectures, seminars, workshops and conferences. The Center cooperates with other universities in Germany, and sponsors visits by scholars from Europe and America.

Director: Manfred Krifka

For further information about ZAS, please consult our website:

<http://www.zas.gwz-berlin.de>

or write to:

Manfred Krifka, Director
Zentrum für Allgemeine Sprachwissenschaft
Schützenstr. 18
D-10117 Berlin
Germany

E-mail: krifka@zas.gwz-berlin.de

ZAS Papers in Linguistics reflect the ongoing work at ZAS. They comprise contributions of ZAS researchers as well as visiting scholars. Issues are available on an exchange basis or on request. For further information, please write to:

Sekretariat
Zentrum für Allgemeine Sprachwissenschaft
Schützenstr. 18
D-10117 Berlin
Germany

E-mail: sprach@zas.gwz-berlin.de
Phone: +49 30 20 19 24 01
Fax: +49 30 20 19 24 02

Later issues can also in part be downloaded at the ZAS website:

http://www.zas.gwz-berlin.de/index.html?publications_zaspil

Cover design: Mathias Krüger, Mechthild Bernhard and the CMS, HU Berlin.

ZAS Papers in Linguistics previous issues (please consult the ZAS website for full tables of content, and for availability):

- ZASPiL 1 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on syntax and semantics. Contributions by Ewald Lang, Anna Cardinaletti & Michal Starke, Jaklin Kornfilt, Ewald Lang, Renate Steinitz and Chris Wilder.
- ZASPiL 2 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on syntax and morphology. Contributions by Peter Ackema & Ad Neeleman, Gaberell Drachman, Ursula Kleinhenz, Sylvia Löhken, André Meinunger, Renate Raffelsiefen, Iggy Roca, M. M. Verhijde and Wolfgang Ullrich Wurzel.
- ZASPiL 3 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on syntax and phonology. Contributions by Ulrike Demske, Damaris Nübling, Wolfgang Sternefeld and Susan Olsen.
- ZASPiL 4 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on syntax and learning. Contributions by Artemis Alexiadou & Elena Anagnostopoulou, Hans-Martin Gärtner, Jaklin Kornfilt, Paul Law, André Meinunger, Ralf Vogel & Markus Steinbach and Chris Wilder.
- ZASPiL 5 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on syntax. Contributions by Artemis Alexiadou & Spyridoula Varlokosta, Elena Herburger, Paul Law, Alan Munn, Cristina Schmitt, Juan Uriagereka, Chris Wilder and Petra de Wit & Maaïke Schoorlemmer.
- ZASPiL 6 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Sylvia Löhken (eds.):
Papers on clitics. Contributions by Artemis Alexiadou & Elena Anagnostopoulou, Piotr Banski, Monika Baumann, Loren A. Billings, Damir Cavar, Uwe Junghanns, Ursula Kleinhenz, Jaklin Kornfilt, Christine Maaßen, Cristina Schmitt, Petra de Wit & Maaïke Schoorlemmer, Maaïke Schoorlemmer, Chris Wilder and Ilse Zimmerman.
- ZASPiL 7 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Ursula Kleinhenz (eds.):
Papers on phonetics and phonology. Contributions by Loren Billings, Christina Kramer & Catherine Rudin, Janet Grijzenhout, T. A. Hall, Haike Jacobs, Peter M. Janker, Manuela Noske, Bernd Pompino-Marschall, Peter M. Janker and Christine Mooshammer.
- ZASPiL 8 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Ursula Kleinhenz (eds.):
Papers on syntax, semantics, phonology and acquisition. Contributions by Artemis Alexiadou & Elena Anagnostopoulou, Artemis Alexiadou & Melita Stavrou, Dagmar Bittner, Hans-Olav Enger, Manuela Friedrich, Wladimir D. Klimonow and Heike Wiese.
- ZASPiL 9 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Ursula Kleinhenz (eds.):
Papers on focus and ellipsis. Contributions by Loren A. Billings, Horst-Dieter Gasde, Uwe Junghanns, André Meinunger, Kerstin Schwabe and Ning Zhang.
- ZASPiL 10 Artemis Alexiadou, Nanna Fuhrop, Paul Law and Ursula Kleinhenz (eds.):
Papers on syntax of clefts, pseudo-clefts, relative clauses, and the semantics of present perfect Contributions by Artemis Alexiadou & Anastasia Giannakidou, Marcel den Dikken, André Meinunger and Chris Wilder, Caroline Heycock & Anthony Kroch, Jason Merchant, Renate Musan, Wolfgang Sternefeld, Peter Svenonius and Chris Wilder.
- ZASPiL 11 Artemis Alexiadou, Nanna Fuhrop, Ursula Kleinhenz and Paul Law (eds.):
Papers on morphology and phonetics. Contributions by H.G. Tillmann, K.J. Kohler, P.A. Keating, F. Schiel & A. Kipp, Ch. Draxler, A. Mengel, R. Benz Müller & M. Grice, A. P. Simpson, L. Ellis & W. J. Hardcastle, K. Russell, E. Farnetani, M. Jessen, B. Kröger, L. Faust and B. Pompino-Marschall & P. M. Janker.
- ZASPiL 12 Artemis Alexiadou, Nanna Fuhrop, Ursula Kleinhenz and Paul Law (eds.):
Papers on morphology and phonology. Contribution by Ursula Kleinhenz.
- ZASPiL 13 Artemis Alexiadou, Nanna Fuhrop, Ursula Kleinhenz and Paul Law (eds.):
Papers on morphology. Contributions by Werner Abraham, Nanna Fuhrop, Livio Gaeta, Rüdiger Harnisch, Heinrich Hettrich, Bernhard Hurch, Wladimir D. Klimonow, Ekkehard König & Peter Siemund, Elisabeth Leiss, Elke Ronneberger-Sibold, Peter Schrijver, Richard Schrodtt, Anja Voeste and Wolfgang Ullrich Wurzel.

- ZASPiL 14 Ewald Lang and Ljudmila Geist (eds.):
Papers on semantics of the copula. Contributions by Ewald Lang, Ljudmila Geist, Claudia Maienborn, Gerhard Jäger, Johannes Dölling, Ilse Zimmermann, Ning Zhang, Renate Musan, Renate Steinitz and Cristina Schmitt.
- ZASPiL 15 Artemis Alexiadou, Nanna Fuhrhop, Ursula Kleinhenz and Paul Law (eds.):
Papers on language change and language acquisition. Contributions by Werner Abraham, Nanna Fuhrhop, Gregory K. Iverson & Joseph C. Salmons, Wladimir Klimonow, Michail Kotin, Peter Suchsland, Letizia Vezzosi, Dagmar Bittner, Manuela Friedrich, Natalia Gagarina, Insa Gülzow and Theodore Marinis.
- ZASPiL 16 Ewald Lang (ed.):
Papers on copular- and AUX-constructions. Contributions by Ewald Lang, Gerhard Jäger, Michail Kotin, Cristina Schmitt, Nanna Fuhrhop, Ljudmila Geist and Joanna Blaszczak
- ZASPiL 17 Cathrine Fabricius-Hansen, Ewald Lang and Claudia Maienborn (eds.):
Approaching the grammar of adjuncts. Proceedings of the Oslo conference. Contributions by Assinja Demijanow & Anatoli Strigin, Johannes Dölling, David Dowty, Thomas Ernst, Marina V. Filipenko, Werner Frey, Graham Katz, Claudia Maienborn, Barbara Partee & Vladimir Borshev, Karin Pittner, Inger Rosengren, Susan Rothstein, Benjamin Shaer, Arnim von Stechow and Ilse Zimmermann.
- ZASPiL 18 Dagmar Bittner, Wolfgang U. Dressler and Marianne Kilani-Schoch (eds.):
First verbs: On the way to mini-paradigms. Contributions by Dagmar Bittner, Wolfgang U. Dressler & Marianne Kilani-Schoch, Sabine Klampfer, Insa Gülzow, Klaus Laalo, Barbara Pfeiler, Marianne Kilani-Schoch, Carmen Aquirre, Antigone Katicic, Pawel Wójcik and Natalia Gagarina.
- ZASPiL 19 T. A. Hall and Marzena Rochon (eds.):
Investigations in prosodic phonology. Contributions by Bozena Cetnarowska, Laura J. Downing, T. A. Hall, David J. Holsinger, Arsalan Kahnemuyipour, Renate Raffelsiefen, Marzena Rochon and Caroline R. Wiltshire.
- ZASPiL 20 Kerstin Schwabe, André Meinunger and Horst-Dieter Gasde (eds.):
Issues on topics. Contributions by André Meinunger, Yen-Hui Audrey Li, Liejiong Xu, Danqing Liu, Marie-Claude Paris, Kleanthes K. Grohmann, Artemis Alexiadou, Werner Frey and Michael Grabski.
- ZASPiL 21 Oliver Teuber and Nanna Fuhrhop (eds.):
Papers for Ewald Lang. Contributions by Dagmar Bittner and Klaus-Michael Köpcke, Werner Frey, Nanna Fuhrhop, Michael Grabski, Kleanthes Grohmann, Tracy Alan Hall, Wladimir D. Klimonov, Paul Law, Kerstin Schwabe, Patrick O. Steinkrüger, Oliver Teuber and Wolfgang Ullrich Wurzel.
- ZASPiL 22 Gerhard Jäger, Anatoli Strigin, Chris Wilder and Ning Zhang (eds.):
Papers on Predicative Constructions. Contributions by John F. Bailyn, Misha Becker Patrick Brandt, Assinja Demijanow & Anatoli Strigin, Roland Hinterhölzl, Orin Percus, Susan Rothstein, Sze-Wing Tang, Wei-Tien Dylan Tsai and Ning Zhang.
- ZASPiL 23 Klaus von Heusinger and Kerstin Schwabe (eds.):
Information Structure and the Referential Status of Linguistic Expressions. Contributions by Franz-Josef d'Avis, Carsten Breul, Dina Brun, Daniel Büring, Donka F. Farkas, Hans-Martin Gärtner, Michael Hegarty, Jeanette K. Gundel & Kaja Borthen, Jüßen Lernerz, Horst Lohnstein, Norberto Moreno & Isabel Pérez, Paul Portner, Ingo Reich, Elisabeth Stark, Anita Steube and Carla Umbach.
- ZASPiL 24 Klaus von Heusinger and Kerstin Schwabe (eds.):
Sentence Type and Specificity. Contributions by Raffaella Zanuttini & Paul Portner, Horst-Dieter Gasde, Kleanthes K. Grohmann, Remus Gergel, Kerstin Schwabe, Klaus von Heusinger, Bart Geurts, Nicholas Asher and Werner Frey.
- ZASPiL 25 Anatoli Strigin and Assinja Demijanow (eds.):
Secondary Predication in Russian. Contributions by Anatoli Strigin and Assinja Demijanow.

- ZASPiL 26 Ning Zhang (ed.):
The Syntax of Predication. Contributions by David Adger & Gillian Ramchand, Tor A. Åfarli & Kristin M. Eide, Ana Ardid-Gumiel, Kleanthes K. Grohmann, Youngjun Jang & Siyoun Kim, Jaume Mateu, Joan Rafel, Kylie Richardson, Peter Svenonius and Ning Zhang.
- ZASPiL 27 Ewald Lang und Ilse Zimmermann (eds.):
Nominalizations. Contributions by Fritz Hamm & Michiel von Lambalgen, Veronika Ehrich, Veronika Ehrich & Irene Rapp, Ulrike Demske, Artemis Alexiadou, Klaus von Heusinger and Ilse Zimmermann.
- ZASPiL 28 T. A. Hall, Bernd Pompino-Marschall and Marzena Rochon (eds.):
Papers on Phonetics and Phonology: The Articulation, Acoustics and Perception of Consonants. Contributions by Hansook Choi, Silke Hamann, Kenneth de Jong, Kyoko Nagao & Byung-jin Lim, Lisa M. Lavoie, Jeff Mielke, Marianne Pouplier & Louis Goldstein, Daniel Recasens, Rachid Ridouane, Zoë Toft, Nathalie Vallée, Louis-Jean Boë, Jean-Luc Schwartz and Pierre Badin & Christian Abry.
- ZASPiL 29 Dagmar Bittner and Natalia Gagarina (eds.):
The Acquisition of Aspect. Contributions by Dagmar Bittner, Annerieke Boland Dina Brun & Babyonyshev, Sophia Delidaki & Spyridoula Varlokosta, Alison Gabriele, Gita Martohardjona & William McClure, Miren Hodgson, Linae Jeschull, Claire Martinot, Maja Andel & Sunil Kumar, Ayumi Matsuo, Barbara Schmiedtová, Yasuhiro Shirai and Ursula Stephany & Maria Voeikove.
- ZASPiL 30 Regine Eckardt (ed.):
Questions and Focus. Contributions by Florian Schwarz and Markus Fischer.
- ZASPiL 31 Dagmar Bittner (ed.):
Von starken Feminina und schwachen Maskulina. Contribution by Dagmar Bittner.
- ZASPiL 32 T. A. Hall and Silke Hamann (eds.):
Papers in Phonology and Phonetics. Contributions by Karen Baertsch, Stuart Davis, Jana Brunner, Susanne Fuchs, Pascal Perrier, Hyeon-Zoo Kim, Antony Dubach Green, T. A. Hall, Silke Hamann, Jaye Padgett and Marzena Zygis.
- ZASPiL 33 Natalia Gagarina and Dagmar Bittner (eds.):
Studies on the Development of Grammar in German, Russian and Bulgarian. Contributions by Dagmar Bittner, Natalia Gagarina, Milena Kühnast, Velka Popova, Dimitar Popov and Franziska Bewer.
- ZASPiL 34 Paul Law (ed.):
Proceedings of AFLA 11, ZAS, Berlin 2004. Contributions by Edith Aldridge, Loren Billings & Daniel Kaufman, Chun-Mei Chen, Wen-yu Chiang & Fang-mei Chiang, Wen-yu Chiang & I Chang-Liao, Mark Donohue, Nelleke Goudswaard, Nikolaus Himmelmann, Arthur Holmer, Arsalan Kahnemuyipour & Diane Massam, Daniel Kaufman, Tomoko Kawamura, Edward Keenan & Cecile Manorohanta, Yuko Otsuka, Ileana Paul, Matt Pearson, Eric Potsdam, Craig Thiersch.
- ZASPiL 35 Ben Shaer, Werner Frey and Claudia Maienborn (eds.):
Proceedings of the Dislocated Elements Workshop, ZAS Berlin, November 2003. Contributions by Maria Alm, Olga Arnaudova, Betty Birner, Ariel Cohen, Cécile de Cat, Judit Gervain, Beáta Gyuris, Liliane Haegeman, Konstantina Haidou, Anke Holler, Ruth Kempson & Ronnie Cann & Jieun Kiaer, Anikó Lipták, Eric Mathieu, Sam Mchombo & Yukiko Morimoto, Nicola Munaro & Cecilia Poletto, Frederick J. Newmeyer, Andreas Nolda, Javier Pérez-Guerra & David Tizón-Couto, Benjamin Shaer & Werner Frey, Nicholas Sobin, Augustin Speyer, Malte Zimmermann.
- ZASPiL 36 Anatoli Strigin:
Blocking Resultative Secondary Predication in Russian.
- ZASPiL 37 Susanne Fuchs and Silke Hamann (eds.):
Papers in Phonetics and Phonology. Contributions by Laura J. Downing, Christian Geng, Antony D. Green, T. A. Hall, Silke Hamann, Al Mtenje, Bernd Pompino-Marschall, Christine Mooshammer, Sabine Zerbian, and Marzena Zygis.

- ZASPiL 38 Jason Mattausch:
On the Optimization and Grammaticalization of Anaphora
- ZASPiL 39 Jana Brunner:
Supralaryngeal mechanisms of the voicing contrast in velars
- ZASPiL 40 Susanne Fuchs, Pascal Perrier and Bernd Pompino-Marschall (eds.):
Speech Production and Perception: Experimental analyses and models. Contributions by Susanne Albert, Jérôme Aubin, Pierre Badin, Sophie Dupont, Sascha Fagel, Roland Frey, Alban Gebler, Cédric Gendrot, Julia Gotto, Abraham Hirschberg, Ian S. Howard, Mark A. Huckvale, Bernd J. Kröger, Ines Lopez, Shinji Maeda, Lucie Ménard, Christiane Neuschaefer-Rube, Xavier Perlorson, Pascal Perrier, Hartmut R. Pfitzinger, Bernd Pompino-Marschall, Nicolas Ruty, Walter Sendlmeier, Willy Serniclaes, Antoine Serrurier, Annemie Van Hirtum and Ralf Winkler.
- ZASPiL 41 Susanne Fuchs:
Articulatory correlates of the voicing contrast in alveolar obstruent production in German.
- ZASPiL 42 Christian Geng, Jana Brunner and Daniel Pape (eds.):
Papers in Phonetics and Phonology. Contributions by Jana Brunner, Katrin Dohlus, Susanne Fuchs, Christian Geng, Silke Hamann, Mariam Hartinger, Phil Hoole, Sabine Koppetsch, Katalin Mády, Victoria Medina, Christine Mooshammer, Pascal Perrier, Uwe D. Reichel, Anke Sennema, Willy Serniclaes, Krisztián Z. Tronka, Hristo Velkov and Marzena Zygis.
- ZASPiL 43 Laura J. Downing, Lutz Marten, Sabine Zerbian (eds.):
Papers in Bantu Grammar and Description. Contributions by Leston Buell, Lisa Cheng, Laura J. Downing, Ahmadi Kipacha, Nancy C. Kula, Lutz Marten, Anna McCormack, Sam Mchombo, Yukiko Morimoto, Derek Nurse, Nhlanhla Thwala, Jenneke van der Wal and Sabine Zerbian.
- ZASPiL 44 Christian Ebert and Cornelia Endriss (eds.):
Proceedings of the Sinn und Bedeutung 10. Contributions by Stavros Assimakopoulos, Maria Averintseva-Klisch, Kata Balogh, Sigrid Beck & Arnim von Stechow, Adrian Brasoveanu, Ariel Cohen, Paul Dekker, Ljudmila Geist, Wilhelm Geuder, Wilhelm Geuder & Matthias Weisgerber, Elsi Kaiser, Elsi Kaiser & Jeffrey T. Runner & Rachel S. Sussman & Michael K. Tanenhaus, Dalina Kallulli, Mana Kobuchi-Philip, Sveta Krasikova & Ventsislav Zhechev, Eric McCready, Telmo Mória, Karina Veronica Molsing, Fabrice Nauze, Francesca Panzeri, Doris Penka, Daniel Rothschild, Florian Schwarz, Torgrim Solstad, Stephanie D. Solt, Tamina Stephenson, Rachel Szekely, Lucia M. Tovená, Anna Verbuk, Matthias Weisgerber, Hedde Zeijlstra, Malte Zimmermann, Eytan Zweig.
- ZASPiL 45 Sabine Zerbian:
Expression of Information Structure in the Bantu Language Northern Sotho
- ZASPiL 46 Ines Fiedler & Anne Schwarz (eds.):
Papers on Information Structure in African Languages. Contributions by Klaus Abels & Peter Muriungi, Enoch O. Aboh, Robert Carlson, Bernard Caron, Klaudia Dombrowsky-Hahn, Wilfrid H. Haacke, Angelika Jakobi, Susie Jones, Gregory Kobele & Harold Torrence, H. Ekkehard Wolff & Doris Löhr.
- ZASPiL 47 Barbara Stiebels (ed.):
Studies in Complement Control
- ZASPiL 48 Dagmar Bittner & Natalia Gagarina (eds.):
Intersentential Pronominal Reference in Child and Adult Language. Proceedings of the Conference on Intersentential Pronominal Reference in Child and Adult Language. Contributions by Jeanette K. Gundel, Dimitris Ntelitheos & Melinda Kowalsky, H. Wind Cowles, Peter Bosch & Carla Umbach, Gerlof Bouma & Holger Hopp, Petra Hendriks, Irene Siekman, Erik-Jan Smits & Jennifer Spenader, Dagmar Bittner, Natalia Gagarina, Milena Kühnast, Insa Gülzow & Natalia Gagarina.

ZASPiL 49 Marzena Zygis & Susanne Fuchs (eds.):

Papers in Phonetics and Phonology. Contributions by Claire Brutel-Vuilmet & Susanne Fuchs, Marzena Zygis, Laura Downing, Elke Kasimir, Daniel Recasens, Silke Hamann & Susanne Fuchs, Anna Bloch-Rozmej, Grzegorz Nawrocki, Cédric Patin.

ZASPiL 50 Hristo Velkov:

Akustische Analysen zur koartikulatorischen Beeinflussung des frikativischen Teils stimmloser Plosive im Deutschen und im Bulgarischen