palgrave
macmillan

Optimality Theory and Pragmatics

Edited by
Reinhard Blutner and Henk Zeevat

Optimality Theory and Pragmatics

Palgrave Studies in Pragmatics, Language and Cognition

Palgrave Studies in Pragmatics, Language and Cognition is a new series of high quality research monographs and edited collections of essays focusing on the human pragmatic capacity and its interaction with natural language semantics and other faculties of mind. A central interest is the interface of pragmatics with the linguistic system(s), with the 'theory of mind' capacity and with other mental reasoning and general problem-solving capacities. Work of a social or cultural anthropological kind will be included if firmly embedded in a cognitive framework. Given the interdisciplinarity of the focal issues, relevant research will come from linguistics, philosophy of language, theoretical and experimental pragmatics, psychology and child development. The series will aim to reflect all kinds of research in the relevant fields – conceptual, analytical and experimental.

Titles include:

Reinhard Blutner and Henk Zeevat (*editors*)
OPTIMALITY THEORY AND PRAGMATICS

Forthcoming titles:

Ira Noveck and Dan Sperber
EXPERIMENTAL PRAGMATICS

Corinne Hen
LINGUISTIC MEANING, TRUTH CONDITIONS AND RELEVANCE

Palgrave Studies in Pragmatics, Language and Cognition Series
Series Standing Order ISBN 0–333–99010–2
(*outside North America only*)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England.

Optimality Theory and Pragmatics

Edited by

Reinhard Blutner and Henk Zeevat

palgrave
macmillan

First published 2004 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N.Y. 10010
Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN 1–4039–0129–5

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Optimality theory and pragmatics/edited by Reinhard Blutner and Henk Zeevat.
p. cm.
Includes bibliographical references and index.
ISBN 1–4039–0129–5 (cloth)
1. Optimality theory (Linguistics). 2. Pragmatics. 3. Grammar, Comparative and general—Syntax. I. Blutner, Reinhard. II. Zeevat, Henk, 1952–

P158.42.O683 2003
415—dc21 2003053267

10 9 8 7 6 5 4 3 2 1
13 12 11 10 09 08 07 06 05 04

Printed and bound in Great Britain by
Antony Rowe Ltd, Chippenham and Eastbourne

To Johanna, Stefan, and Robert

This page intentionally left blank

Contents

Acknowledgments

This book is the result of a series of conferences, workshops and informal meetings that aimed to bring together Optimality Theory and theories of natural language interpretation. The series of discussions were opened at "The First Conference on the Optimization of Interpretation" held in Utrecht, January 4–5, 2000, organized by Petra Hendriks, Helen de Hoop, Fabien Reniers and Frank Wijnen. Less formal meetings were held in Szklarska Poreba (Poland), where the "Workshops on the Roots of Pragmasemantics" (March 18–21, 2001, and March 2–6, 2002) gave us an opportunity for intensive midnight discussions about some of the main ideas of the present book. We are especially grateful to the staff of the Tourist Hostel on the Szrenica Peak for providing such a beautiful and relaxed atmosphere, and to our local organizer, Anna Pilatova, for also managing the most dangerous situations.

In the final phase of the book project, a workshop on "Optimality Theory and Pragmatics" (Berlin, ZAS, June 8–10, 2002) was of essential importance for directing our efforts and for preparing the chapters in the present book. Special thanks go to Anatoli Strigin who not only did the local organization, but also made funds available that allowed us to invite our contributors from the other side of the ocean. Further, we are deeply indebted to our referees who helped very much in improving the manuscripts. Thanks to Paul Boersma, Daniel Büring, Elena Karagjosova, Elena Maslova, Marie Nilsenova, Roger Schwarzschild and Carla Umbach. This book has been written by its authors and we feel that in merely saying thank you to them all, we do not sufficiently express our appreciation of their own contributions.

Finally, we owe a great debt to Robyn Carston who proposed that we write this book and gave valuable advice and support.

Reinhard Blutner and Henk Zeevat

Notes on the Contributors

David Beaver is a faculty member at Stanford University. He is based in the Linguistics Department and is an affiliate of the Symbolic Systems Program. Beaver's work has been concerned with formal semantics and pragmatics of natural language. He is the author of *Presupposition and Assertion in Dynamic Semantics* (CSLI Publications) and of numerous papers about pragmatic subjects.
(email: dib@stanford.edu)

Reinhard Blutner is *Privatdozent* at the Humboldt-University in Berlin. He began his scientific career in theoretical physics and shifted later to artificial intelligence, cognitive psychology and theoretical linguistics. In his work he integrates insights from connectionist psychology, logic, computer science and cognitive linguistics. Currently, he is a lecturer in Artificial Intelligence and Cognitive Philosophy at the University of Amsterdam.
(email: blutner@hum.uva.nl)

Helen de Hoop is Assistant Professor in general linguistics at the University of Nijmegen, the Netherlands. Her 1992 Ph.D. thesis on case configuration and noun phrase interpretation appeared in 1996 as a book in the Garland series Outstanding Dissertations in Linguistics. Lately, she has contributed to the development of optimization methods in semantics by organizing several international conferences on this topic and publishing articles in this area, co-authored with Petra Hendriks, Henriëtte de Swart and Jaap van der Does.
(email: H.deHoop@let.kun.nl)

Hans-Martin Gärtner received his Ph.D. from the University of Frankfurt/Main in 1997 with a thesis on minimalist syntax. He is co-editor of a volume on *The Role of Economy Principles in Linguistic Theory* (1997). Currently, he is Assistant Director at ZAS, Berlin.
(email: gaertner@zas.gwz-berlin.de)

Petra Hendriks received her Ph.D. in 1995 from the University of Groningen. She wrote her thesis on comparative constructions in categorial grammar. Currently, she is Assistant Professor at the departments of Dutch and Artificial Intelligence at the University of Groningen.
(email: P.Hendriks@ppsw.rug.nl)

Gerhard Jäger received his Ph.D. in 1996 from Humboldt-University in Berlin with a thesis on formal semantics. After holding various research

positions in Munich, Philadelphia, Berlin and Utrecht, he is now *Privatdozent* at the Computational Linguistics Department of Potsdam University. In addition, he conducts a research project on bidirectional Optimality Theory at the ZAS Berlin.
(email: jaeger@zas.gwz-berlin.de)

Hanjung Lee received her Ph.D. from Stanford University with a thesis on the optimization in argument expression and interpretation. She is currently at the University of Minnesota. She is interested in research in the morphosyntax of East Asian and Southwest Asian languages, syn-tactic typology and variation, Optimality Theory, corpus linguistics and psycholinguistics.
(email: hanjung-lee@hotmail.com)

Jason Mattausch is a research assistant at the 'Zentrum für allgemeine Sprachwissenschaft' in Berlin. His current research interests involve the application of insights from formal semantics, computational linguistics and pragmatics to theories of discourse production and interpretation and Optimality Theoretic approaches to syntax and semantics.
(email: mattausch@zas.gwz-berlin.de)

Ralf Vogel is research assistant at the Institute of Linguistics, University of Potsdam. His area of specialization is the syntax of the Germanic languages, and, more generally, the interaction between syntax, semantics and phonology, and how it can be modeled within Optimality Theory. His former research and teaching stations were at the Humboldt University Berlin and the University of Stuttgart.
(email: rvogel@ling.uni-potsdam.de)

Jennifer Spenader received her B.A. in linguistics from the University of Illinois in 1992. She obtained her Ph.D. degree at Stockholm University in the Department of Linguistics with a thesis on presuppositions in spoken discourse. Her other interests include abstract object anaphora and computational approaches to discourse.
(email: jennifer@ling.su.se)

Robert van Rooy is a KNAW Fellow working on the project: Games, Relevance and Meaning. His research area includes formal semantics and pragmatics of natural language, and philosophy of language. He is stationed at the Department of Philosophy at the University of Amsterdam.
(email: vanrooy@hum.uva.nl)

Henk Zeevat is Senior Lecturer at the Department of Computational Linguistics and at the ILLC at the University of Amsterdam. He contributed to the development of discourse representation theory, unification grammar and Optimality Theoretic semantics. His research and teaching positions were at the Universities of Rotterdam, Edinburgh and Stuttgart.
(email: henk.zeevat@hum.uva.nl)

This page intentionally left blank

1

Editors' Introduction: Pragmatics in Optimality Theory

Reinhard Blutner and Henk Zeevat

Based on the tenets of the so-called 'radical pragmatics' school (see, for instance, Cole, 1981), this book takes a particular view with regard to the relationship between content and linguistically encoded meaning. The traditional view embodied in the work of Montague and Kaplan (e.g., Kaplan, 1979; Montague, 1970) sees content being fully determined by linguistic meaning relative to a contextual index. In contrast, the radical view takes it that, although linguistic meaning is clearly important to content, it does not determine it, as pragmatic principles also play a role. The central issue of this book is how to give a principled account of the determination of content. Seeing linguistic meanings as underdetermining the content (proposition) expressed, there must be a pragmatic mechanism of completion which can be best represented as an optimization procedure. It is demonstrated that the general framework of Optimality Theory (OT) makes it possible to formulate the desired explanatory principles.

The first section of this general introduction outlines the basic framework of OT as applied to phonology, syntax and morphology. The second section takes a historical perspective and shows that the idea of optimization was present in the pragmatic enterprise right from the beginning. Further, it explains the main advantages of the general framework of OT when applied to the field of pragmatics, and it puts the whole idea into concrete terms by demonstrating how Horn's (1984) theory of conversational implicature can be implemented within a bidirectional optimality theory. In Section 3, we raise several basic questions underlying the whole volume and discuss them from a theoretical and empirical perspective. This part gives an overview of the different topics treated in the book, and it explains in which respects the single contributions aim to satisfy our cooperative goal: to give a new impulse to the tradition of radical pragmatics. Section 4, finally, outlines basic open questions of future research.

1 Optimality Theory

OT was initiated by Prince and Smolensky (1993) as a new phonological framework that deals with the interaction of violable constraints. In recent years, OT was also the subject of lively interest outside phonology. Students of morphology, syntax and natural language interpretation became sensitive to the opportunities and challenges of the new framework. The reasons for this growing interest in OT are empirical and conceptual. First, it turned out that a series of empirical generalizations and observed phenomena can be expressed very naturally within this framework; this holds especially for phonology where in-depth analyses of many languages have provided a much better insight into cross-linguistic tendencies than we had before the invention of OT. Second, and perhaps much more important in linking scientists into a new research paradigm, there are the conceptual reasons, which are many in the present case: (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 problem of learning hidden structure and the logical problem of language acquisition; and (v) the aim to integrate the synchronic with the diachronic view of language.

OT respects the generative legacy in two important methodological aspects: the strong emphasis on formal precision in grammatical analysis and the goal of restricting the descriptive power of linguistic theory. Seeing themselves within the Generative tradition, many representatives of OT adopt the fundamental distinction between Universal Grammar (UG) and a language-specific part of Grammar. UG describes the innate knowledge of language that is shared by all normal humans, and aims both to describe the universal properties of language and the range of variation possible among languages. The language-specific part of grammar typically consists of the lexicon and a system reflecting the specific structural properties of the particular language. Within the generative tradition, the concrete theoretical realization of this distinction has changed over the years. In the principles and parameters model, for example, UG is conceptualized as a system of (inviolable) principles, which are parameterized to demarcate the space of possible forms (see, for instance, Chomsky, 1981). The fixing of these parameters (triggered by language-specific data) determines the grammar of the particular language. OT realizes an essentially different view of this distinction.

At this point we must emphasize that optimality theory is rooted, at least in part, in connectionism, a paradigm that makes use of neurobiological assumptions – in an extremely simplified way. As a consequence, OT does not assume a strict distinction between representation and processing. More than ten years ago, there was a lively debate in cognitive linguistics concerning the true architecture of cognition – the debate between connectionists and symbolists. The proponents of a symbolic architecture, among them Fodor and Pylyshyn (e.g., Fodor and Pylyshyn, 1988), had the clever idea of taking the arguments for connectionism as showing that symbolic architecture is *implemented* in a certain kind of connectionist network. This idea corresponds to the strategy of maintaining classical architecture and reducing connectionism to an implementation issue. The development of OT demonstrates that the opposite strategy is more exciting: augmenting and modifying symbolist architecture by integrating insights from connectionism.

Let's take a closer look now at the background and the nature of OT. Like other models of grammars, OT sees a grammar as specifying a function that assigns to each input (underlying representation of some kind) a structural description or output. For example, in Grimshaw and Samek-Lodovici's theory of the distribution of clausal subjects (e.g., Grimshaw and Samek-Lodovici, 1998), an input is a lexical head with a mapping of its argument structure into other lexical heads, plus a tense specification. The input also specifies which arguments are foci and which arguments are coreferent with the topic. An example is:

(1) <sing(x), x = topic, x = he; T = pres perf>

It represents the predicate *sing*, with a pronominal argument that is the current discourse topic. A possible output is an X-bar structure realizing an extended projection of the lexical head. Examples are:

(2) a. [$_{\text{IP}}$ has [sung]] a clause with no subject
b. [$_{\text{IP}}$ he$_{\text{i}}$ has [t$_{\text{i}}$ sung]] a clause with subject *he*, co-indexed with a trace in SpecVP
c. [$_{\text{IP}}$ has [t$_{\text{i}}$ sung]he$_{\text{i}}$] *he* right-adjoined to VP, co-indexed with a trace in SpecVP

The general idea of standard versions of generative syntax is to define the acceptable (grammatical) input–output pairs via a system of rules and transformations. In order to restrict the descriptive power of linguistic theory, constraints are added. All of these constraints have been viewed as inviolable within the relevant domain. The idea of inviolable constraints has itself proved to be problematic and this has led to the "parametrization" of certain constraints, with one parametric setting for one language and a different parametric setting for another language.

In OT the "generative part" of the grammar is reduced to a universal function *Gen* that, given any input *I*, generates the set *Gen*(*I*) of candidate structural descriptions for *I*. The central idea of OT is to give up the inviolability of constraints and to consider a set *Con* of violable constraints. Furthermore, a strict ranking relation >> is defined on *Con*. This relation makes it possible to evaluate the candidate structural descriptions in terms of the totality of the violations they commit, as determined by the ranking of the constraints. If one constraint C_1 outranks certain constraints $C_2, \ldots, C_i$, written $C_1 \gg \{C_2, \ldots, C_i\}$, then *one* violation of C_1 counts more than as arbitrarily many violations of $C_2, \ldots, C_i$. The evaluation component selects the optimal (least offending, most harmonic) candidate(s) from the set *Gen*(*I*). The grammar favors the competitor that best satisfies the constraints. Only an optimal output is taken as an appropriate (grammatical) output; all suboptimal outputs are taken as ungrammatical. This idea makes the grammaticality of a linguistic object dependent on the existence of a competitor that better satisfies the constraints.

Constraints are of two different kinds: *markedness constraints* that affect outputs only and *faithfulness constraints* that relate to the similarity between input and output. The main representatives of the faithfulness family are: (i) Parse prohibiting *underparsing* ("underlying input material is parsed into output structure") and (ii) Fill prohibiting *overparsing* ("the elements of the output must be linked with correspondents in the underlying input"). In OT-syntax the latter constraint is also called Full-Int(erpretation): the elements of the output must be interpreted. Markedness constraints are inherently connected with the domain under discussion. By way of example, we consider the following two constraints in the case of our OT syntax (distribution of clausal subjects):

(3) a. Subj "the highest A-specifier in an extended projection must be filled"[1]
    b. Drop-Topic "arguments coreferent with the topic are structurally unrealized"

To complete this short introduction to OT, let's consider a typical OT tableau relating to the input–output pairings (1) and (2). In the present example, the following constraint hierarchy is assumed:

(4) Full-Int >> Drop-Topic >> Parse >> Subj

As can be seen from tableau (5), this ranking yields an "Italian" behavior in which topicalized subjects are suppressed; this is exemplified by the optimal parse (5a).

(5)

<sing(x), x = topic, x = he; T = pres perf>	Full-Int	Drop-Topic	Parse	Subj
☞ (a) [$_{IP}$ has [sung]]			*	*
(b) [$_{IP}$ he$_i$ has [t$_i$ sung]]		*		
(c) [$_{IP}$ has [t$_i$ sung] he$_i$]		*		*

This behavior would change to "English" if we chose the following hierarchy:

(6) Parse >> Subj >> Full-Int >> Drop-Topic

where Parse and Subj outrank Drop-Topic. In this case, (5b) would arise as the optimal candidate.

The architecture of OT suggests a simple realization of the fundamental distinction between UG on the one hand and the language-specific part of Grammar on the other hand: UG consists of *Gen* (the generator) and *Con* (the set of constraints); the language-particular aspect of Grammar is determined by the particular ranking of the constraints. This proposal bolsters the way for defining a factorial typology:

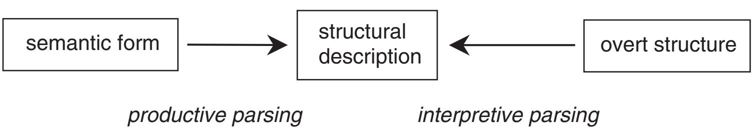
> *Typology by reranking*: Systematic crosslinguistic variation is due entirely to variation in language-specific total rankings of the universal constraints in *Con*. Analysis of the optimal forms arising from all possible total rankings of *Con* gives the typology of possible human languages. UG may impose restrictions on the possible rankings of *Con*.
>
> (Tesar and Smolensky, 2000, p. 27)

As already shown in Prince and Smolensky (1993), analysis of all rankings of the constraints considered in the basic CV syllable theory reveals a typology that explains Jacobson's (1962) typological generalizations. In the case of OT-syntax, Grimshaw and Samek-Lodovici (1995) were the first who performed an analysis involving all rankings of the above constraints and derived a typology of subject distribution in this way.

Typology by reranking is the most famous but not the only pleasant consequence from the general architecture of OT. Another consequence is the idea of *robust interpretive parsing*, which is substantial for many purposes, such as psycholinguistic applications of OT in describing online language production, comprehension and natural language acquisition.

Although the term *parsing* is used more commonly in the context of language comprehension, in the OT literature it is treated as the general issue of assigning structure to input, an issue relevant to both comprehension and production. To be sure, the canonical perspective of an OT grammar is related to production – taking the input as an underlying form, and the output structural description as including the surface form. This type of parsing is called "productive parsing", and it is schematically represented in diagram (7) – the term "overt structure" is used instead of "surface structure":

(7)



In the context of language comprehension, another mapping comes into play. It maps a given overt form to an optimal structural description *SD* whose overt portion matches the given form. The process of computing the optimal *SD* for an overt form is called *interpretive parsing*.

It is a common observation that competent speakers can often construct an interpretation for utterances they simultaneously judge to be ungrammatical. Whereas it is notoriously difficult to account for this kind of "robustness" of natural language interpretation within rule- or principle-based models of language, the interpretation of ungrammatical sentences is much simpler when using an OT architecture. *Robust interpretive parsing* is the idea of parsing an overt structure with a grammar even when that structure is not grammatical according to that grammar. It is important to recognize that the presence of interpretable, but ungrammatical sentences immediately corresponds to mismatches between productive and interpretive parsing. Consider an interpretive parse that starts with some overt structure *OS* and assigns an optimal structural description *SD*. Paired with *SD* is a certain semantic form *SF*. The grammaticality of *SD* (and its overt structure, *OS*) depends on whether the outcome of productive parsing leads us back to *SD*, when starting with *SF*. In case it does, then *SD* is grammatical; otherwise, it is ungrammatical.

As a simple illustration we reconsider the earlier example from OT syntax. Let's take the constraint hierarchy (4) that accounts for "Italian" syntactic behavior. In Italian, sentences such as *he has sung* are unacceptable if the pronoun refers to a discourse topic. Using the hierarchy (4), this is demonstrated in tableau (5), where the sentence *he has sung* comes out as suboptimal. Despite its unacceptability, the sentence is parsed into a structural description, namely $[_{\text{IP}}$ he$_{\text{i}}$ has [t$_{\text{i}}$ sung]]. An important point in all examples of this kind is that both in productive parsing and in interpretive parsing the same constraint hierarchies are used. The difference arises solely from the different candidate sets that are relevant for the different perspectives of optimization.

The idea of robust interpretive parsing is crucial for the mechanism of language learning in OT when it is combined with another idea – the idea of constraint demotion (cf. Tesar and Smolensky, 2000). The latter idea conforms to a mechanism that reranks the constraints in a particular way, such that one prearranged candidate becomes the winner over the rest of the candidates (cf. **Vogel**[2]). The combination of both ideas gives the following picture of children's language acquisitions. Becoming confronted with some overt datum, the child tries to understand this datum (on the basis of her current grammar). She performs interpretive parsing, resulting in a structural description that includes an underlying structure. Next, the child turns to the production perspective: she starts with the underlying form and performs productive parsing. If the results of productive and interpretive parsing are different, then this information is used to correct the grammar.

The child applies constraint demotion taking the interpretive parse to be the winner (correct analysis) and the productive parse to be the loser. The child has succeeded in learning the target grammar if interpretive and productive parsing always give the same structural descriptions. Note that an overt form will allow the learner to improve his grammar just in case the current grammar (incorrectly) declares it to be ungrammatical.

There is an important consequence of this view of learning. The OT learning algorithm establishes an interesting type of equilibrium: *what we produce we are able to understand adequately and what we understand we are able to produce adequately*. This equilibrium corresponds to a strong conception of bidirectional optimization: a logical combination of optimal comprehension and optimal generation (cf. Blutner, 2000; Zeevat, 2000; **Beaver and Lee**). Hence, bidirectional optimality can be seen as a kind of synchronic law describing the results of language learning. It should be mentioned that Tesar and Smolensky's (2000) mechanism for learning hidden structures is one aspect of language learning only. Acquiring conventions that link structured forms and conceptual contents (via lexical entries and idiom chunks) is another aspect. Most interestingly, empirical investigations have shown that also in this case the general pattern of bidirectionality or symmetry seems to apply.[3] In the present volume, **Jäger** explores this possibility of bidirectional learning within an evolutionary setting.

Before we apply OT to the domain of pragmatics we must clarify what the general conditions are that every OT system has to satisfy. The following three conditions are the core of OT. They are a necessary basis for the family of procedures that performs grammar learning in OT (Tesar and Smolensky, 2000):

(A) Universal Grammar is assumed to be determined by a generative part *Gen* and a system of violable constraints *Con* (UG = *Gen* + *Con*). The language-specific part of Grammar relates to a particular ranking of the constraints in *Con*. Only this part of the Grammar is learnable. Language learning simply reduces to inferring the ranking of the constraints in *Con*. This excludes both the possibility that the constraints themselves are learned (in part at least) or that aspects of the generator are learnable. On the other side, it excludes the possibility that the set of the possible rankings is constrained on a universal basis.

(B) The force of strict domination $\gg$: A relation of the form $C \gg C'$ does not merely mean that the cost of violating C is higher than that of violating C′; rather, it means that no number of C′ violations is worth a single C violation. The force of strict domination excludes cumulative effects where many violations of lower ranked constraints may overpower higher ranked constraints.

(C) The OT grammar of the language that has to be learned is based on a *total* ranking of all the constraints: $C_1 \gg C_2 \gg \ldots \gg C_n$. This condition

> is crucial for the convergence of the proposed learning mechanism (Tesar and Smolensky, 2000). It can be shown that the iterative procedure of constraint demotion converges to a set of totally ranked constraint hierarchies in this case, each of them accounting for the learning data.

What is the status of these conditions? One way to look at these conditions is to see them as oversimplifications that are made mainly for didactic and practical reasons. Oversimplifications may be needed to allow one to concentrate on a central problem and to sweep aside many problems that are less critical for understanding the central one (i.e., the problem of learning 'hidden' structure.) Moreover, oversimplifications may be necessary to achieve interesting mathematical results that simply are not possible without them. But it is not necessary to see them as simplifications, we can also see them as conditions reflecting the true nature of the domain under discussion and thus are taken to be empirically justified.

It must be admitted that it is not always simple to find out which position really is taken by the representatives of OT. For example, concerning the condition (C), we find the following statement in Tesar and Smolensky (2000):

> From the learnability perspective, the formal results given for Constraint Demotion depend critically on the assumption that the target language is given by a totally ranked hierarchy. This is a consequence of a principle implicit in Constraint Demotion. This principle states that the learner should assume that the description is optimal for the corresponding input, and that it is the *only* optimal description. This principle resembles other proposed learning principles, such as Clark's Principle of Contrast and Wexler's Uniqueness Principle.
>
> (p. 47 ff.)
>
> It appears likely to us that learning languages that do not derive from a totally ranked hierarchy is in general much more difficult than the totally ranked case. If this is indeed true, demands of learnability could ultimately explain a fundamental principle of OT: UG admits only (adult) grammars defined by totally ranked hierarchies.
>
> (p. 50)

Taking condition (C) as a kind of principle that indicates when language learning is simple, however, is a different idea than taking it as a strict demand on theories of learning. In our opinion, the first idea is right and the second wrong. There are many examples where the target language produces synonymies (scrambling data in German and Korean may provide a case in point). We agree that this can delay learning in one case or the other. In this vein, the suggestion is to take (C) as a kind of oversimplification, the acceptance of which is justified only for doing the first significant research steps. Notably, Anttila and Fong (2000) take a similar view (cf. also **Beaver and Lee**). As a consequence, the condition (C) should be given up in an advanced stage and a more general theory should be developed, a theory that *explains* (C) as a principle about the complexity of language learning. In our opinion, Paul Boersma's learning theory (Boersma, 1998; Boersma and Hayes, 2001) is on the right track for doing this job.

With regard to the condition (B), Smolensky himself sees it as a "regimentation and pushing to extremes of the basic notion of Harmonic Grammar" (Prince and Smolensky, 1993, p. 200). And Gibson and Broihier (1998) argue that this restriction does not appropriately characterize the manner in which parsing preferences interact.

What about condition (A)? Many representatives of OT seem to consider it as a *conditio sine qua non*. Boersma's work on functional phonology (Boersma, 1998), however, puts forward convincing arguments exposing principle (A) likewise as a kind of oversimplification.

These questions about the status of the conditions (A)–(C) becomes highly relevant when we try to extend the domain of applications for OT, especially when we try to apply the OT framework to the domain of pragmatics. Hence, for pragmatics in OT, debating and clarifying the status of the condition (A)–(C) is an opportunity and challenge. Most chapters in this volume are directly or indirectly concerned with this task.

2 Pragmatics in OT

The idea of optimization was present in the pragmatic enterprise from the very beginning. Much more than in other linguistic fields, optimality scenarios are present in most lines of thinking: Zipf's (1949) balancing between effect and effort; the Gricean conversational maxims (Grice, 1975, 1989); Ducrot's argumentative view of language use (e.g., Ducrot, 1980); the principle of optimal relevance in relevance theory (Sperber and Wilson, 1986/1995). However, in the course of the development of OT, the area of OT semantics and pragmatics was developed after everything else. This appears rather puzzling, and the reasons for it are not very clear. There may be stylistic aspects that might frighten a serious semanticist or logician: the curious tables with shadows, and the famous little hands. A more serious reason may have to do with an unfortunate 'dynamic turn' which was directed against Kamp's (1981) programmatic outline of a cognitively oriented approach to language. In contrast to Kamp's original paper, which is based on the tenets of 'radical pragmatics', much research which falls under the rubric of the 'dynamic turn' is in the spirit of the conservative view of language which radical pragmatics sets itself against.

While the compositionality assumption underlying the 'dynamic turn' has strengthened the methodology of semantics, it has also led to a mechanistic approach at points where pragmatics and semantics are difficult to keep apart. The habit of interpreting trees with fully resolved pronouns fails to make the distinction between rule-based grammar and the complicated salience weighting of different antecedents required for pronoun resolution, a process that leads to preferences at best. In treatments of presupposition, sometimes presupposition boils down to a single logical operator and so obscures the distinction between the semantic role of presuppositions for their triggers and their function as a sign that the speaker is making an assumption, a distinction that also shows up as the distinction between treating a presupposition by resolving it to the context or by accommodating it. We would submit that an OT theory has an advantage over logical or grammatical treatments in that the ideal of rational cooperative communication can be almost directly captured by constraints that directly derive from Grice's analysis of this cooperative behavior (cf. **van Rooy; Zeevat**).

What can be called with more justice 'radical pragmatics' (cf. Cole, 1981) is to hypothesize a division of labor between: (i) a linguistic system determining the semantic representation of a sentence (Grammar including the lexicon) and (ii) a pragmatic system constituting the interpretation of the corresponding utterance in a given setting (contextual information, encyclopedia). The pragmatic system is taken as realizing Grice's (1975) idea of conversational implicature, and it is modeled with the instruments of OT. As a consequence, many linguistic phenomena which had previously been viewed as belonging to the semantic subsystem, in fact can be explained within the pragmatic subsystem of OT.

Before we enter the discussion concerning in which way optimality theory may help to close the gap between formal (linguistic) meaning and interpretation, we have to consider this distinction more closely. For Grice (1975) the theoretical distinction between what the speaker explicitly said and what he has merely implicated is of particular importance. What has been said is supposed to be based purely on the conventional meaning of a sentence, and is the subject of semantics. What is implicitly conveyed (scalar and conversational implicatures) belongs to the realm of pragmatics. It is assumed to be calculable on the basis of the setting – a notion already introduced by Katz and Fodor (1963), and referring to previous discourse, sociophysical factors and any other use of "non-linguistic" knowledge. Fruitful as this theoretical division of labor may have been – especially as a demarcation of the task of logical semantics – it has inherent problems. More often than not, what is said by a speaker's use of a sentence already depends on the context. Even for Griceans, propositional content is not fully fleshed out until reference, tense and other indexical elements are fixed. However, propositional content must be *inferred* in many cases – going beyond the simple mechanism of fixing indexical elements.

Proponents of relevance theory (see, for example, Carston, 2002, 2003a, 2003b; Sperber and Wilson, 1986) have pointed out that the pragmatic reasoning used to compute implicated meaning must also be invoked to fill out underspecified propositions where the formal meaning contributed by the linguistic expression itself is insufficient to give a proper account of truth-conditional content. A similar point was made in lexical pragmatics (e.g., Blutner, 1998, 2002). Both relevance theory and lexical pragmatics agree in assuming a Gricean mechanism of pragmatic strengthening in order to fill the gap between formal, linguistic meaning and the propositional content (i.e., the *explicit* assumptions communicated by an utterance – called explicature in relevance theory; cf. Sperber and Wilson, 1986/1995, p. 182).

In a similar vein, de Hoop and de Swart (2000) and Hendriks and de Hoop (2001) argue that, with regard to the theory of interpretation, what compositional semantics gives us is a radically underspecied notion of meaning represented by a possibly infinite set of interpretations of a well-formed syntactic structure. In addition, these authors were the first to propose using the framework of optimality theory in order to select the optimal interpretation associated with a particular syntactic structure. For that purpose, they propose a particular set of constraints and rankings between those constraints, based on general principles of rational communication. The interpretive perspective on optimization provides insights into different phenomena of interpretation, such as the determination of quantificational structure and domain restriction (Hendriks and de Hoop, 2001), nominal and temporal anaphora (de Hoop and de Swart, 2000), and the interpretational effects of scrambling (de Hoop, 2000).

Stimulated by Horn's (1984) theory of conversational implicature and related ideas in relevance theory, Blutner (2000) argued that this design of OT is inappropriate and too weak in a number of cases. This is due to the fact that the abstract generative mechanism (*Gen*) can pair *different* forms with one and the same interpretation. The existence of such alternative forms may lead to *blocking* effects which strongly affect what is selected as the preferred interpretation. The phenomenon of blocking has been demonstrated in a number of examples where the appropriate use of a given expression formed by a relatively productive process is restricted by the existence of a more "lexicalized" alternative to this expression. One case in point was provided by Householder (1971). The adjective *pale* can be combined with a great many color words: *pale green, pale blue, pale yellow*. However, the combination *pale red* is limited in a way that the other combinations are not. For some speakers *pale red* is simply anomalous, and for others it picks up whatever part of the pale domain of red *pink* has not preempted. This suggests that the combinability of *pale* is fully or partially blocked by the lexical alternative *pink*.

The phenomenon of blocking requires us to take into consideration what else the speaker could have said. As a consequence, we have to go from a one-dimensional, to a two-dimensional (bidirectional) search for optimality.[4] As mentioned in Section 1, bidirectional optimality can be seen as describing the equilibrium that results from language learning at its limits.

In the domain of pragmatics, the bidirectional view was independently motivated by a reduction of Grice's maxims of conversation to two principles: the Q-principle and the I-principle (Atlas and Levinson, 1981; Horn, 1984, who writes R instead of I). The I/R-principle can be seen as the *force of unification* minimizing the speaker's effort, and the Q-principle can be seen as the *force of diversification* minimizing the hearer's effort (cf. Horn, 1984). The Q-principle corresponds to the first part of Grice's quantity maxim (*make your contribution as informative as required*), while it can be argued that the countervailing I/R-principle collects the second part of the quantity maxim (*do not make your contribution more informative than is required*), the maxim of relation and possibly all the manner maxims. Conversational implicatures which are derivable essentially by appeal to the Q-principle are called Q-based implicatures. Standard examples are scalar implicatures and clausal implicatures. I-based implicatures, derivable essentially by appeal to the I-principle, can be generally characterized as enriching what is said via inference to a rich, stereotypical interpretation (cf. Atlas and Levinson, 1981; Gazdar, 1979; Horn, 1984; Levinson, 2000).

In a slightly different formulation, the I/R-principle seeks to select the most coherent interpretation, and the Q-principle acts as a blocking mechanism which blocks all the outputs which can be grasped more economically by an alternative linguistic input (Blutner, 1998). This formulation makes it quite clear that the Gricean framework can be conceived of as a bidirectional optimality framework which integrates expressive and interpretive optimality. Whereas the I/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. The important feature of this formulation within bidirectional OT is that although it compares alternative syntactic inputs with one another, it still helps to select the optimal meaning among the various possible interpretational outputs of the single actual syntactic input given, by acting as a blocking mechanism.

The so-called strong version of bidirectional OT – it conforms to the equilibrium established during OT learning – can be formulated as given in (8). Here, pairs $\langle f, m\rangle$ of possible (syntactic) forms f and utterance meanings (= interpretations) m are related by means of an ordering relation $<$, *being less costly* (*more harmonic*). At the moment, the precise metric underlying this ordering relation is still open, and the sign $<$ is not much more than a place holder for such a metric. In OT, the ordering relation $<$ can be constituted by a system of ranked constraints, as discussed in many contributions to this volume. Another option would be to work with a single, graded markedness constraint such as RELEVANCE (see **van Rooy**).

(8) **Bidirectional OT** (*Strong Version*)

A form-meaning pair (*f, m*) is optimal iff it is realized by *Gen* and it satisfies both the I- and the Q-principle, where:

a. (f, m) satisfies the I-principle iff there is no other pair (f, m′) realized by *Gen* such that (f, m′) < (f, m)
b. (f, m) satisfies the Q-principle iff there is no other pair (f′, m) realized by *Gen* such that (f′, m) < (f, m)

It should be mentioned that the I-principle is very much in line with the mono-directional view on optimality theoretic interpretation as proposed by de Hoop and de Swart (2000) and Hendriks and de Hoop (2001), which exclusively adopts the hearer's perspective on disambiguation. What is interesting in (8) is that it also implements the Q-principle, thereby also taking the speaker's perspective into account. Hence, a proper treatment of interpretation in OT has to take into account both the perspective of the hearer and the perspective of the speaker. Because this framework of bidirectional OT can be characterized in game-theoretical terms (Dekker and van Rooy, 2000), optimality theoretic pragmatics can be given a proper formal interpretation.

One of the main advantages of the optimality theoretic framework is that it allows the isolation of three substantial components of the overall mechanism: (i) the generator, which provides the potential form interpretation pairs; (ii) the underlying metric, possibly constituted by a system of ranked constraints; and (iii) the two perspectives of optimization. In relevance theory it is *relevance* that constitutes the underlying metric; in other frameworks' notions of information, efficiency and salience are more important (cf. **van Rooy**).

There are, however, several old problems with assuming full symmetric bidirectionality to phonological and syntactic processing in both directions. In phonology, the problem is mostly discussed as the *Rad/Rat* problem (cf. Hale and Reiss, 1998). The German word *Rat* (council) is pronounced as [rat] without any change from the underlying form to the surface form. The word *Rad* (wheel) is pronounced in the same way but here two constraints come into play: the DEVOICING constraint that prefers the pronunciation [rat] to [rad] and FAITHFULNESS that would prefer the pronunciation [rad] and is outranked by DEVOICING in German. If we want to apply the same constraints in the direction from pronunciation to optimal underlying form, *Rat* is always preferred because of FAITHFULNESS in interpretation. The same problem can arise in syntactic ambiguities. Again in German, the sentence *Welches Mädchen mag Reinhard?* is ambiguous between *Which girl likes Reinhard?* and *Which girl does Reinhard like?* The Wh-object has a longer road to go from its canonical position to its sentence initial position than the corresponding Wh-subject. The constraint Stay (= *Do not move*) adopted by most OT syntacticians then prefers the reading with the Wh-subject. Since there is general agreement that there is a proper ambiguity in these cases, full bidirectionality needs to be restricted by some principle which makes the system less symmetric than the Tesar and Smolensky-learning algorithm assumes. In this volume, **Jäger** uses an asymmetric bidirectional system for his learning algorithm, **Vogel** restricts his OT-syntax by powerful pragmatic principles and **Beaver and Lee** consider different ways to avoid the *Rat/Rad* problem in their survey of bidirectionality.

Another problem has to do with the specific features of blocking we find in natural languages. The scenario of strong bidirection describes the case of **total** blocking where some forms (e.g., **furiosity*, **fallacity*) do not exist because others do *(fury, fallacy)*. However, blocking is not always total but may be **partial**, in that only those interpretations of a form are ruled out that are preempted by a "cheaper" competing form. 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 restricted by the existence of a corresponding lexical causative.

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

Whereas lexical causatives – for example, (9a) – tend to be restricted in their distribution to the stereotypic causative situation (direct, unmediated causation through physical action), productive (periphrastic) causatives tend to pick up more marked situations of mediated, indirect causation. For example, (9b) could be used appropriately when Black Bart caused the sheriff's gun to backfire by stuffing it with cotton. 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, p. 26) – a tendency that Horn terms the *division of pragmatic labor*.

There are two principal possibilities for avoiding the fatal consequences of total blocking that are described by strong bidirection. The first possibility is to make some stipulations concerning ***GEN*** in order to exclude equivalent semantic forms. The second possibility is to weaken the notion of (strong) optimality in a way that allows us to derive Horn's division of pragmatic labor in a principled way by means of a sophisticated optimization procedure.

In Blutner (1998, 2000) it is argued that the second option is much more practicable and theoretically interesting. A recursive variant of bidirectional optimization was proposed (called *weak* bidirection) which was subsequently simplified by Jäger (2002):

(10) **Bidirectional OT** (*Weak Version*)
A form-meaning pair (*f, m*) is called super-optimal iff (*f, m*) $\in$ *Gen* … and:

a. there is no other super-optimal pair (f, m′) : (f, m′) < (f, m)
b. there is no other super-optimal pair (f′, m) : (f′, m) < (f, m)

Under the assumption that < is transitive and well-founded, Jäger (2002) proved that (10) is a sound recursive definition and is equivalent to the formulation in Blutner (1998, 2000). In addition, he proved that each pair which is *optimal* (strong bidirection) is *super-optimal* (weak bidirection) as well, but not vice versa. Hence, weak bidirection gives us a chance to find additional super-optimal solutions. For example, weak bidirection allows marked expressions to have an optimal interpretation, although both the expression and the situations they describe have a more efficient counterpart. Hence, this formulation is able to describe Horn's division of pragmatic labor. The notion of weak bidirection is discussed in more detail by **Mattausch** (Chapter 4, Section 3.2).

The existence of two notions of bidirectionality raises a conceptual problem: which conception of bidirectionality is valid, the strong or the weak one? Obviously, this question relates to the foundation of bidirection in an overall framework of cognitive theory. As we have already seen, the *strong mode* of optimization in (8) – *what we produce we are able to understand adequately and what we understand we are able to produce adequately* – corresponds to the equilibrium established by the OT-learning algorithm. Hence, the strong conception of bidirectionality can be seen as a kind of synchronic law describing the results of language learning.

Weak bidirection gives a chance of finding additional solutions. Is it possible to give a natural interpretation for these additional solutions? We want to propose the idea that these additional solutions are due to the ability and flexibility of self-organization in language change which the weak formulation alluded to. In other words, we propose to take these additional solutions as describing the possible outcomes of self-organization before the learning mechanism has fully realized the equilibrium between productive and interpretive optimization.

Jäger (2002) and Dekker and van Rooy (2000) have proposed algorithms that update the ordering (preference) relation < such that (i) optimal pairs are preserved and (ii) a new optimal pair is produced if and only if the same pair was super-optimal at earlier stages. Consequently, we can take the solutions of weak bidirection to be identical with the solutions of strong bidirection considering all the systems that result from updating the ordering relation. Recently, van Rooy (forthcoming) and **Jäger** (this volume) have reconsidered this problem and have proposed algorithms within an evolutionary setting – realizing a mechanism of self-organization in language change. This point may be clarified when we (re)consider Horn's *division of pragmatic labor* and relate it to the principle of *constructional iconicity* in the school of "natural morphology" (for references see Wurzel, 1998):

> **Constructional iconicity**: A semantically more complex, derived morphological form is unmarked regarding constructional iconicity, if it is symbolized formally more costly than its semantically less complex base form; it is the more marked, the stronger its symbolization deviates from this.
>
> (Wurzel, 1998, p. 68)

In this school the principle plays an important role in describing the *direction of language change*. In fact, constructional iconicity and Horn's division of pragmatic labor can be proven to be a consequence of weak bidirection. This observation gives substance to the claim that *weak bidirection* can be considered as a principle describing (in part) the direction of language change: super-optimal pairs are tentatively realized in language change. This relates to the view of Horn (1984) who considers the Q-principle and the I-principle as diametrically opposed forces in inference strategies of language change. Of course, the idea goes back to Zipf (1949), and is reconsidered in van Rooy (forthcoming).[5] Arguing that Horn's division of pragmatic labor is a *conventional* fact about language, this convention can be explained in terms of equilibriums of signaling games introduced by Lewis (1969) – making use of an evolutionary setting (see van Rooy, forthcoming).

But is it really the case that weak bidirectionality does not play a role in synchrony? The Horn example is the pair *Black Bart shot the sheriff/ Black Bart caused the sheriff to die*. A similar example is Grice's *Mrs T produced a series of sounds closely resembling the score of "Home Sweet Home"*, which contrasts with: *Mrs T. sang "Home Sweet Home"*. Horn's and Grice's point is that the long and unusual form are used to convey that there was something special with the killing and the singing and that this is not accidental. The process by which this special interpretation is arrived at cannot be diachronic language change: the long and unusual forms are so unusual that it is not possible to assume a special conventionalization process that associates the special meaning with the special form.

Grice's explanation from his maxim *Be Brief* can be almost directly translated in OT pragmatics. The relevant constraint is Economy, which we can reinterpret as the requirement that there is no correct form interpretation pair that is more economical (or more standard?) in either dimension. This immediately leads us to reject the association of the complex (unusual) forms with the standard meaning: for that we have a simpler and more usual form. It likewise rules out the association of the simple form with the non-standard meaning. The result is that we obtain an underspecified special meaning for the special forms which must be interpreted further with respect to the context and the situation to give us the concrete interpretations (kill in a bizarre way, sing rather badly) that we seem to obtain. Notice, however, that the speaker has not said any of this, she has merely suggested that there is something special going on. There is no convention that fixes the meaning. The vagueness and cancellability of the extra interpretation suggests that we are dealing with an implicature and not with part of the truth-conditional content.

There are three points to be made about this reinterpretation of Grice's stylistic maxim. In the first place, it is a very low constraint which can be overridden by any grammatical or semantical constraint that one needs to assume. It is the lowest of the low. Second, it is obviously weakly bidirectional for it to work. If the standard-form/marked-meaning or the marked-form/standard-meaning were in competition with marked-form/marked-meaning, that last pair would not survive. And third, it seems that – with some charity – all other pragmatic principles can be related to it. As Blutner and Jäger show (1999), the constraint DO NOT ACCOMMODATE can be seen as a special case of semantic economy, minimizing the number of new discourse referents. The constraint RELEVANCE can also be seen as a kind of semantic economy: irrelevant information is information that the interlocutor is not seeking for and requires the accommodation of new questions or interests of the interlocutor. Information that is consistent or consistent with the context is pragmatically less complex than information that is inconsistent in itself or inconsistent with the context. The whole of pragmatics would be weakly bidirectional under this interpretation.

If this were the case, it would also give us an indication of why weak bidirectionality is such a powerful explanatory principle in diachronic linguistics. Pragmatic weak bidirectionality creates special interpretations that can become conventionalized. Assume that a marked form is used with some frequency to indicate the same marked meaning. It will then become a conventional device to indicate the marked meaning, and the marked meaning will no longer be derived by weak bidirectionality but by a lexical or grammatical convention. Think about Hebrew optional object case marking conventionally meaning that the referent is definite. Or about the Dutch *wijf* – originally the standard word for woman, but pushed away by *vrouw* (originally mistress) – that can now only be used for the purpose of expressing contempt for the referent in question.

Summarizing, we suggest taking the strong conception of bidirectionality as a synchronic law and the weak one as conforming to diachrony (with the reservation and clarification just sketched). In addition, the present conception conforms to the idea that synchronic structure is significantly informed by diachronic forces. Further, it respects Zeevat's (2000) acute criticism against super-optimality as describing an online mechanism (see also **Beaver and Lee**). From the perspective of *grammaticalization*, we are very close to Hyman's (1984) dictum of seeing grammaticalization as the harnessing of pragmatics by a grammar. And there are connections to a recent proposals by Haspelmath (1999) for an OT-based theory of language change.

3 Overview

The aim of this book is to demonstrate that OT also finds fruitful applications in the domain of pragmatics and can contribute in overcoming the gap between linguistic meaning and utterance meaning. This section contains an overview of the different topics treated in the book and it explains in which respects the single contributions aim to satisfy our cooperative goal: giving the tradition of radical pragmatics a new impulse.

The promise of OT pragmatics is that by using the OT architecture, some order can be brought to the seemingly unrelated approaches that constitute pragmatics. There have been a series of studies that try to reformulate treatments of pragmatic phenomena to optimality theory. De Hoop and de Swart (2000) study the determination of quantifier restrictions, a classical challenge to compositional semantics, since that determination is only partially determined by the syntactic tree, and can involve interactions with the context, the information structure and the linear order of the quantifier. One of the factors in the solution is relating the interpretation to given material, either in the topic or in the context. This problem area comes back in studying pronoun syntax and resolution (Beaver, to appear; Bresnan, 2001), presupposition (Jäger and Blutner, 2000; Zeevat, 2000), the binding theory (Burzio, 1991, 1998; Levinson, 1987a, 2000). Other areas of pragmatics where OT has been attempted are intonation and information structure (Beaver and Clark, 2002; Schwarzschild, 1999), scalar implicatures (Blutner, 2000; van Rooy, 2001).

In the present volume, Helen **de Hoop** (Chapter 2) provides an in-depth discussion based on real data of the Complementary Preference Hypothesis as an account of stressed pronouns in English and formulates an alternative account in terms of two interpretive constraints: Contrastive Stress and Continuing Topic, to overcome the problems with the earlier account.

Petra **Hendriks** (Chapter 3) combines a semantic analysis of only (*only(A)(B) = all(B)(A)*) with an OT account of how intonation and syntax conspire in determining the scope and restrictor of determiners and focus-sensitive particles. The account builds on earlier work of de Hoop and de Swart (2000) and Hendriks and de Hoop (2001) using optimality theoretic semantics.

Jason **Mattausch** (Chapter 4) introduces the influential work of Levinson on the origin and typology of binding theory and reformulates the different historical stages assumed by Levinson in bidirectional optimality theory. The reformulation is able to avoid and solve a number of problems in Levinson's proposal and can avoid the M-principle altogether, which comes out as a theorem in bidirectional optimality theory.

Henk **Zeevat** (Chapter 5) reviews an earlier attempt to treat 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, and so on.

A proper framework of OT is also the correct platform for asking foundational questions. Given that we have violable principles and reliability ranking between them (let's assume this can be decided on empirical grounds), what follows about the representations on which the constraints have to work, can a rational foundation be found for each of the constraints and can the order between the constraints be founded in some rational principle? The notions of relevance and economy have particularly been in focus here. Another foundational issue concerns the nature of bidirection and the symmetry assumption (e.g., Zeevat, 2000). Further questions concern the division of labor between semantics and pragmatics in particular, and the modularity stipulation in general. And what is the proper architecture of an overall system integrating elements from syntax, prosody, semantics and pragmatics?

In the present volume, David **Beaver** and Hanjung **Lee** (Chapter 6) give an overview of various proposals in bidirectional optimality theory where crucial tests are total, along with partial, blocking, the *Rat/Rad* problem and some other problems. They show conclusively that weak super-optimality cannot be combined with standard proposals for optimality theoretic syntax with a larger number of constraints.

Gärtner's analysis in Chapter 7 of Icelandic object-shift and differential marking of (in-)definites in Tagalog addresses the issue of disambiguation in natural languages. In the first part he suggests a family of OT-constraints called "Unambiguous Encoding", which can be understood as a correlate of Gricean "Avoid Ambiguity". In the second part he points out some shortcomings of this approach, and he suggests that the OT-status of "Unambiguous Encoding" is epiphenomenal. Two ways of reduction are explored which bolster the way for a functionalist understanding of the phenomenon – viewing grammars as "harnessed" or "frozen" pragmatics (cf. Hyman, 1984). In addition, and not unrelated to the contribution of **Beaver and Lee**, he points out some serious problems for Blutner's version of bidirectional OT.

Robert **van Rooy** argues in Chapter 8 that the general framework of optimality theoretic pragmatics is able to include basic insights from relevance theory. Starting from the bidirectionality of Blutner (1998, 2000) in terms of the Q- and I-principles, he develops a decision-theoretic notion of relevance to take – in the first instance – the place of the Q-principle in this scheme for pragmatics. Though this leads to improvements, further problems then force the tentative adoption of a relevance-based exhaustivity operator as a basis for reconstructing the Q-principle, the I-principle and Blutner's bidirectionality. Horn's M-principle is then derived by minimization of effort.

Ralf **Vogel** in Chapter 9 addresses the problem of OT architecture. Following Jackendoff (1997) he assumes three levels of representation: a semantic (= conceptual), a syntactic and a phonological level. The *correspondence* between these levels is modeled by a (bi-directional) OT grammar. Arguing that syntax is much less encapsulated and 'autonomous' than generative grammar usually assumes, Vogel's model is able to restrict OT-syntax by powerful pragmatic principles. In addition, there is a methodological point that deserves particular attention. The proposed architecture is not only motivated by its ability to account for certain intriguing linguistic phenomena. It is also justified by its compatibility with current OT learning theory.

OT pragmatics is a theory of pragmatic competence that invites the crossing of boundaries in traditional pragmatics and of relating it to psycholinguistic theories of natural language performance (both production and comprehension) on the one hand, and to theories of language learning and language evolution on the other. This volume contains two contributions that explicitly conform to this challenge.

Jennifer **Spenader**'s psycholinguistic investigation in Chapter 10 concerns the choice between two demonstrative forms in Swedish (one simple, the other compound). A multitude of factors influence the choice of one referential form over another, such as abstractness, animacy, and the level of activation of the referent. The general finding is that the simple form is typically used with more accessible and salient referents, while the compound form is used for referents with a lower level of activation. Spenader argues that stochastic optimality theory is capable of modeling the subtle, yet statistically significant differences between the two demonstrative forms – making use of constraints that are independently motivated.

The contribution of Gerhard **Jäger** in Chapter 11 can be seen as the first step in a long research agenda which derives from the view that many syntactic and semantic facts are frozen pragmatics. It should be possible to show how particular languages emerge from pragmatics assuming the fairly standard account of the evolution of phonological forms. Even the advantages involved in moving from a purely pragmatic language to a language with partial conventionalization can be studied from this perspective. The potential contribution of OT here is twofold. OT can inspire learning algorithms and it can provide the framework for the representation and evolution of grammatical knowledge. The diachronic perspective here offers a far more sophisticated picture of the mode of existence of a language. It is not just a conventional association between form and meaning, happening on some rather poorly understood hardware and offering a window on the nature of that hardware, it is one of the possible conventional associations that has a certain degree of stability due to the conditions under which language is transferred to ever-new speakers, their ways of organizing these data, and the frequencies with which the various elements making up the association are used.

In particular, **Jäger** applies a bidirectional generalization of Boersma and Hayes's (2001) learning algorithm to the formalization and simulation of the grammaticalization processes underlying case systems. He is able to show that structural case is the natural outcome of pragmatic case marking, and that some systems are stable whereas others are either unlearnable or very unstable. The account also explains and underpins Aissen's (2000) treatment of differential case marking.

4 Problems and perspectives

The OT approaches to pragmatic phenomena seem to gain empirical advantages with respect to their non-OT predecessors, but that is not the only advantage. Important is the fact that we gain a different way of talking about these things in which uniformities can appear across the description of the different phenomena and that we have the prospect of a single theory of pragmatics where all the phenomena come together. This unification is still a prospect but there are a number of issues that can already be distinguished.

The first issue is the existence of a pragmatic factorial typology. If there is a factorial typology, then it would fly in the face of the pragmatic tradition that has always maintained that pragmatics is universal and consists of a few principles that can be founded in the conceptual analysis of linguistic communication, as in Stalnaker (1999), Grice (1989), Sperber and Wilson (1986/1995), Levinson (1983, 2000), Horn (1984, 2003) and others.

Is it really possible that a constraint CONSISTENT (sometimes treated as part of GEN) could be outranked by a constraint like ECONOMY OF EXPRESSION or RELEVANCE? This would mean that there could be communities where it is more important to be economical than to be consistent with the context, or more important to be relevant than to be consistent with the context. In the first case, it would not be possible to mark corrections; in the second, the interpretation process would maximize relevance without bothering about what we know already. It seems, though, that there is a general functional case for keeping corrections apart from consistent updates since the changes that have to be made to the knowledge of the interpreter are quite different. Rerankings of this kind have to our knowledge not been found in the language communities of the world or only marginally (e.g., politeness can override sincerity).

If we succeed in agreeing on a universal system of ranked pragmatic constraints, there arises a second issue – a foundational one. It concerns the need not just for explaining why there are these constraints and no others and why they are ranked in this way. Because of the lack of variation, the factorial typology does not help to support an empirical argument that our system is correct. Possible strategies are the classical one of deriving the pragmatic system from pure reason, other strategies might try to use an evolutionary argument, which establishes that the pragmatic system is an evolutionary stable state by showing that any mutations (rerankings, small changes to the individual constraints) are eliminated and moreover that it is the only evolutionary stable state among a range of competitors. Of our contributors, **van Rooy** and **Jäger** are following these different strategies, and it is one of the important questions of future research how to relate these different approaches (see van Rooy, forthcoming, for a first step in answers from this direction).

The third issue is how to reconcile universal pragmatics with the obvious fact that there is a great deal of variation in the syntactic, lexical and phonological expression of pragmatic properties in the languages of the world. It is an important insight that even if we have a pragmatic system, this does not mean that pragmatics is purely universal. Languages exhibit enormous differences in their inventory of pragmatically relevant items, like in pronouns (for a basic typology, see Bresnan, 2001), tense and aspect, definite and indefinite markers, presupposition triggers, elliptical constructions, discourse particles. They also differ widely in their marking strategies for information structure. The richness of the data here is still largely unexplored especially in their interaction with the pragmatic treatments that have been the focus of OT pragmatics. It is unclear to what extent these typological variations reflect on the abstract semantics. In Bresnan (2001), we see that Chichewa free pronouns (i.e., the closest analogon to English pronouns) do not allow antecedents that are topical, unlike English, where the pronoun predominantly refers to topical elements. The difference is that, in Chichewa, there is a class of bound pronouns realized in verbal agreement morphology that are used whenever the antecedent is considered to be a topic. Chichewa is not so different from French: French clitic pronouns are used for topic, the free pronouns are used for the other cases. (These cases are not so easy to delineate.) The morphological distinction between zero, bound, clitic and free pronouns is not realized in all languages, but seems to align in different ways with a prominence hierarchy on the antecedents. Whether this hierarchy is universal cannot be decided on the current state of research. The hierarchy itself may be universal, but it is clear from data in Gundel, Hedberg and Zacharski (1993) that zero pronouns do not align with the same property in the different languages that have them in their inventory.

For example, it cannot be decided yet whether pronoun resolution can be split into a part to be treated in OT syntax and general pragmatic constraints on pronoun resolution. If one follows **van Rooy**, the general principle is relevance. It would seem that for a particular treatment of, for example, Chichewa free pronouns, resolution would need additional facts about the Chichewa inventory and the preference of bound pronouns for topical discourse referents.

A fourth issue is the nature of pragmatic constraints. One of the special features of the constraints that seem useful in pragmatics is that they seem like small OT competitions on their own. Consider a neutral, and as far as we know original, example that is reasonably well understood, the resolution of ellipsis:

(11) Jan heeft een rode wollen trui gekocht en Piet drie blauwe.
"Jan has a red woolen sweater bought and Piet three blue"

The resolution process maximizes the similarity between the antecedent sentence and the ellipsed sentence. In a syntactic copying perspective, it copies the verb, the auxiliary, the object noun and one of the object adjectives. It does not copy the color adjective, the subject and the object determiner. It is clear that higher order unification, a tree assimilation algorithm, computation of the most specific common denominator, and source reconstruction – to mention only some of the techniques that have been applied to ellipsis – all attempt to make the ellipsed sentence as similar as possible to its antecedent. This can be naturally described as an OT competition.[6] The point is that constraint violations to a constraint MAXIMIZE SIMILARITY must be scored by the existence of more similar candidates and that there is no alternative to that, since correctness of the resulting sentence misses out on the presence of optional material in the antecedent sentence, predicting, for example, that (12) is a correct interpretation even though the adjective *wollen* ("woolen") is not taken along.

(12) Piet heeft drie blauwe truien gekocht.
"Piet bought three blue sweaters"

This seems the correct way to score DO NOT ACCOMMODATE, ECONOMY and RELEVANCE and STRENGTH, the main pragmatic constraints that people have come up with. We see whether there are otherwise correct interpretations with less discourse referents, otherwise correct sentences with less nodes and words that have the same interpretation in the context, or interpretations that deal with more questions that the interlocutor can be assumed to entertain.

In concluding these introductory remarks, we want to stress once more that OT gives us a powerful instrument for implementing basic pragmatic mechanisms. However, one should not forget that having a hammer in one's hands may seduce one into seeing everything as a nail. For that reason, methodological considerations for restricting the proper domain of OT applications in the area of pragmatics are important, and the significance of the three general conditions (A)–(C) of Section 1 deserves special attention in the area of pragmatics.[7] On the other hand, we are at the beginning of a deeper understanding of our instrument, which – unlike a real hammer – has proven to be helpful in quite different respects. Possibly, it will facilitate the integration of syntax, prosody and pragmatics. It may allow the development of an evolutionary perspective showing that particular language traits emerge from pragmatics. And it may well provide a new research framework in psycholinguistics.

With any luck, the present volume helps to give a start.

Notes

1. Roughly, this condition states that the subject position must not be empty.
2. In this Introduction, names in bold type without a date refer to contributions to this volume.
3. See, for instance, Hayes and Hayes (1989) and Green (1990). Studies with chimpanzees have shown that they typically fail the symmetry test, but children older than two years pass it (Dugdale and Lowe, 2000).

   It should be noticed that the first half of the equilibrium's condition – *what we produce we are able to understand adequately* – follows from the assumed initial state of the OT Grammar (the markedness constraints outrank the faithfulness constraint) plus the assumed mechanism of constraint demotion. In contrast, the second half of the condition – *what we understand we are able to produce adequately* – is independent of the initial state and an immediate consequence of the learning mechanism. In the more general case of learning arbitrary codes, it needs extra requisites to ensure the symmetry condition. For example, it requires a particular asymmetry between expressive and productive optimization (see Zeevat, 2000; **Jäger**).
4. The origin of these ideas goes back to Blutner, Leßmöllmann, and van der Sandt (1996) and Blutner (1998).
5. A very similar point was made in functionalist phonology (e.g., Boersma, 1998). Most 'phonetically driven' or functionalist theories of phonology propose that two of the fundamental forces shaping phonology are the need to minimize effort on the part of the speaker and the need to minimize the likelihood of confusion on the part of the listener. The need to avoid confusion is hypothesized to derive from the communicative function of language. Successful communication depends on listeners being able to recover what a speaker is saying. Therefore it is important to avoid perceptually confusable realizations of distinct categories; in particular, distinct words should not be perceptually confusable. The phonology of a language regulates the differences that can minimally distinguish words, so one of the desiderata for a phonology is that it should not allow these minimal differences, or contrasts, to be too perceptually subtle. There is nothing new about the broad outlines of this theory and it very closely relates to Zipf's (1949) two opposing economies (see also Lindblom, 1986, 1990; Martinet, 1955).
6. A competition with different flavors of resolution arising from different data structures that have to be made as similar as possible, and the possibility of having different maxima to account for ambiguities. This is not the place to take a stance on the empirical and computational issues involved here.
7. Concerning the condition B, for instance, an interesting and new hypothesis is that the hierarchical encoding of constraint strengths is correlated with the effect of automaticity in psychological processes. Perhaps it is the area of pragmatics where this hypothesis can be tested in the most effective way.

2
On the Interpretation of Stressed Pronouns

Helen de Hoop

In the 1974 movie *The Conversation* the utterance *He'd kill us if he got the chance* plays a major role.[1] The leading actor, Gene Hackman, tape-records a conversation made by a couple for a client. The 'he' refers to Hackman's client and the 'us' refers to the couple. Hackman's immediate interpretation of the recorded utterance is that his client might actually kill the conversation participants, that is, the 'us'. But in the final part of the movie, contrary to Hackman's expectation, it is the couple who kill Hackman's client, and not the other way around. After this surprising outcome, we hear the central utterance from the recording once more. However, now we hear it as *He'd kill US if he got the chance*, with stress on 'us'. Did we previously misunderstand the utterance or did we miss the stress on the plural pronoun? No, the director has deliberately manipulated the recording and in doing so, radically changed our interpretation of the utterance. The final time the recording is played, 'us' is pronounced differently, thereby affecting the entire plot. Only at that point in the movie do we understand the recorded message as actually communicating the couple's intention to kill Hackman's client; when they say *He'd kill US if he got the chance*, they actually mean: *We'd better kill him (before he kills us).*

1 Stress on anaphoric pronouns

All languages in the world appear to have personal pronouns, but they come in different forms, for instance full versus reduced ones or free versus bound ones. In languages that have both reduced and non-reduced pronouns, the reduced ones are specialized for anaphoricity, the non-reduced ones have focus functions (cf. Bresnan, 2001). In languages that do not have different types of pronouns, the interaction with prosody gives the same result: unstressed pronouns need less effort; hence, they are specialized for anaphoricity, while the stressed ones have focus functions. However, stress is used for different reasons in language (new information, contrast, shift in reference) and it is not always clear what principles guide a hearer's interpretation of a stressed or unstressed pronoun in a certain context. What are the different types of constraints that play a part in (un)stressed pronoun resolution and how do these interact? In this chapter the interpretation of stressed pronouns in discourse will be analysed in an optimality theoretic fashion.

Pronouns are usually studied in their anaphoric uses, although it is well known that they can be used deictically as well. For example, Bosch (1983) gives the following question–answer pair:

(1) Did anybody leave that lecture yesterday?

(2) HE left.

As Bosch notes, in reply to the question in (1) there would be nobody in the focus of attention who *he* could link up to. Instead, *he* would have to bring somebody into focus that has not been in focus already. In such a case, *he* is accented, indicated by the capitals in (2). Bosch (1983) adds "in order to bring into focus someone who has not already been the focus of attention, *he*, in the deictic use, would most naturally be accompanied by a pointing gesture" (Bosch, p. 58). In written texts, this pointing is sometimes described, as in the following examples (boldface is mine; pronouns that were put in italics by the author to indicate that they are stressed, are replaced by pronouns in capitals):

(3) In their fort on the Lynx Hills the three Lynkestids, the sons of Aioropos, stood on their brown stone ramparts. It was an open place, safe from eavesdroppers. **They had left their guest downstairs**, having heard what he had to say, but given no answer yet. Around them stretched a rear sky of white towering clouds, fringed with mountains. It was late spring; on the bare peaks above the forests, only the deepest gullies showed veins of snow.
'Say what you like, both of you', said the eldest, Alexandros, 'but I don't trust it. What if this comes from the old fox himself, to test us? Or to trap us, have you thought of that?'
'Why should he?' asked the second brother, Heromenes. 'And why now?'
'Where are your wits? He is taking his army into Asia, and you ask why now.'
'Well,' said the youngest, Arrabaios, 'that's enough for him surely, without stirring up the west? No, if it had been that, it would have come two years ago, when he was planning to march south.'
'As HE says' – **Heromenes jerked his head towards the stairway** – 'now's the time. Once Philip's set out, he will have his hostage for us.'
He looked at Alexandros, whose feudal duty it was to lead their tribal

levies in the King's war. He stared back resentfully; already before this, he had been thinking that once his back was turned, the others would ride out on some mad foray that would cost him his head.[2]

[HE = the guest downstairs]

(4) 'This lad's only nineteen', said Heromenes. 'If Philip dies now, with no other son besides the lackwit, then YOU' – **he stabbed his finger at Alexandros** – 'are next in line'.

[YOU = Alexandros]

In most cases, however, pronouns are anaphoric. Anaphoric pronouns refer to individuals already introduced and salient in the discourse. There is an antecedent in the linguistic context to which the pronoun is anaphorically linked. As such, anaphoric pronouns are often continuing topics or at least, they are part of the background and not in focus. Because of this, anaphoric pronouns are usually deaccented, yet this is not necessarily the case. Anaphoric pronouns may be stressed as well, in which case the accent does not indicate deixis, nor the introduction of a novel referent in the discourse, but rather it signals *contrast* in the discourse. Consider the following examples of dialogues from Vallduví (1990) (stress indicated by capitals again):

(5) S1: Good morning. I am here to see Mrs Bush again.
S2: Sure, Mr Smith. Let's see ... One of her assistents will be with you in a second.
S1: Could I see HER today? I'm always talking to her assistents.

(6) [At a grocery's cash register]
S1: It's $1.20 ... o.k. ... Here's your change and here's your broccoli.
S2: Thank you.
S1: Thank YOU.

In the dialogue in (5), the stressed pronoun *her* is anaphoric as it refers back to Mrs Bush. Constituent (or narrow) focus evokes contrast within a contextually salient set of alternatives (Rooth, 1992). A rhetorical relation of contrast is established between two similar but in at least one respect crucially different events (Mann and Thompson, 1988; Asher, 1993). In the dialogue in (5) the relevant contrast is between the event of seeing one of Mrs Bush's assistants versus the event of seeing Mrs Bush herself. That is, in (5) the conversational implicature evoked by the accent on *her* is that the first speaker *does not* want to see one of Mrs Bush's assistants (cf. Rooth, 1992). Similarly, in (6), the accent on *you* in *thank you* establishes a relation of contrast between the event when speaker 1 thanks speaker 2 and the one when speaker 2 thanks speaker 1. That is, the conversational implicature is understood as: *Don't thank me*.

Consider as a further illustration of the contrast evoking function of stress the following paradigm:

(7) a. HE kissed me.
    b. He kissed ME.
    c. HE kissed ME.

Without further context, the pronoun *he* in (7a) can be interpreted either deictically or anaphorically. When it's an anaphoric pronoun, the stress evokes a contrastive interpretation that may be paraphrased as *HE, not somebody else, kissed me*. Similarly, (7b) is paraphrased as *He kissed ME, not somebody else*. Now, what happens in (7c)? I claim that the relation of contrast evoked by the two stressed pronouns in (7c) – in the absence of further context – is interpreted as the contrast between two similar yet crucially different situations, namely one when he kissed me and the other when I kissed him. The implicature of *HE kissed ME* can thus be formulated as *And not the other way around*.[3]

2 Kameyama's complementary preference hypothesis

In this chapter I will argue against Kameyama (1999), who claims to have a unified account of interpretation preferences of stressed and unstressed pronouns in discourse. Kameyama's central intuition is expressed as the "Complementary Preference Hypothesis", taking the interpretation preference of the unstressed pronoun as the base from which to predict the interpretation preference of the stressed pronoun in the same discourse position.

(8) *Complementary Preference Hypothesis (CPH)*: A focused pronoun takes the complementary preference of the unstressed counterpart.
(Kameyama, 1999, p. 315)

So, Kameyama claims that the preferred value of a stressed anaphoric pronoun in discourse is predictable from the preferred value of its unstressed counterpart, and that they draw their values from the same 'currently salient' subset of the domain. The problem of choosing among alternative values for pronouns has been investigated in the framework of centering theory (Grosz, Joshi and Weinstein, 1995). Unstressed pronouns, in particular, are primarily used to indicate the *backward-looking center*, or as I will call it in this chapter, the *continuing topic*. In Kameyama's approach, an unstressed pronoun normally realizes a 'maximally salient entity' of an appropriate number–person type. This, for example, accounts for the preference for a pronoun to corefer with the matrix subject in the previous utterance as in the following example, discussed by Kameyama (1999):

(9) John hit Bill. Mary told him to go home.
[him = John]

In (10), however, world knowledge about the relation *hit* (namely, that when *x* hits *y, y* is normally hurt) overrules the fact that John is more salient than Bill, which results in Bill preferred over John for the unstressed counterpart of *he*.[4] As a consequence, the complementary preference hypothesis makes John preferred over Bill for the stressed pronoun in (11):

(10) John hit Bill. Then he was injured.
[he = Bill]

(11) John hit Bill. Then HE was injured.
[HE = John]

Thus, Kameyama's Complementary Preference Hypothesis correctly derives the right interpretation for the stressed pronoun in (11).

Kameyama, furthermore, discusses the following two famous sequences (cf. Lakoff, 1971):

(12) Paul called Jim a Republican. Then he insulted him.
[Paul insulted Jim]

(13) Paul called Jim a Republican. Then HE insulted HIM.
[Jim insulted Paul]

On the basis of these examples, Kameyama claims there to be a systematic relation between the stressed and unstressed counterparts, which is of a complementary preference within a suitable subset of the domain. The assumption is that stressed and unstressed counterparts choose their values from the same salient subset of the domain of individuals.

(14) Jack and Mary are good friends. {He/HE} is from Louisiana.
[He/HE = Jack]

In other words, in (14) the Complementary Preference Hypothesis cannot be applied, which would make it unclear why stress would be used. Kameyama (1999) argues that when the salient subset is a singleton, as in (14), the focus constraint for the stressed pronoun is satisfied by accommodation. For (14) this means that a contrasting presupposition *Mary is not from Louisiana* is accommodated.

In the following, I will argue against Kameyama's analysis of the use of stressed pronouns in (11), (13) and (14). On the basis of several counterarguments, I will reject the Complementary Preference Hypothesis.

3 Contrast

I would like to claim that the preferred interpretation of all the stressed pronouns above, is in fact the *contrastive* reading. In (11), repeated below as (15), this makes sense precisely in view of our world knowledge about *hit*.

(15) John hit Bill. Then HE was injured.
[HE = John]

The contrast evoked by the stressed pronoun is between the unexpected situation when John is injured as the result of his hitting somebody else and the 'normal' situation when Bill is injured as a result of being hit. Thus, we get the interpretation *Then JOHN was injured* with the implicature *and not Bill (contrary to what you might expect)*. It is not a coincidence that stress is used in a context where a relation of contrast is easily evoked by the sequence of predicates that is used: *hit – being injured*. Additional evidence for the natural occurrence of stress in the example in (15) is that it is maintained if we replace the pronoun by *John*: *Then JOHN was injured*.

Beaver (to appear) uses a different example to illustrate the Complementary Preference Hypothesis, where in my opinion the judgements and the interpretation of stress is far less clear than in Kameyama's example (15). The fragment discussed by Beaver is given in (16):

(16) Fred was eating. He saw Jim. HE winked.

According to Beaver, the stressed pronoun *HE* is interpreted as *Jim*, in accordance with Kameyama's Complementary Preference Hypothesis. However, I have some problems with the interpretation of the stressed pronoun in (16). I guess I would like to claim that the pronoun is still ambiguous as long as the stress is not naturally interpreted as signaling contrast between two events, simply because in our world knowledge there is no obvious connection between either *to see* or *to be seen* and *to wink*. So, the implicature evoked by the stress (*Somebody else did not wink*) is not by itself interpretable without further context, which makes the sequence in (16) harder to interpret than Kameyama's (15). This intuition is supported by the observation that stress is maintained when we replace the pronoun by a proper name in (15), but not in (16). That is, there is no tendency at all to stress the second occurrence of Jim in *Fred was eating. He saw Jim. Jim winked*. In fact, stressing *Jim* here (*Fred was eating. He saw Jim. JIM winked*) sounds odd, just as *HE winked* sounds odd in this context, in my view.

For Kameyama, the preference order among alternative values for the stressed pronoun in (15) (John, Bill) is the complement of the preference order for its unstressed counterpart (Bill, John). However, in (17) below, *John* and *Mary* cannot be alternative values for the same pronoun. Yet, I would like to claim that in (17) we get the same type of reading for the stressed pronoun. That is, the stressed pronoun again evokes a rhetorical relation of contrast between two similar yet crucially different situations. Clearly, the Complementary Preference Hypothesis cannot account for that effect, as Mary does not provide an alternative value for *HE*:

(17) John hit Mary. Then HE was injured.
[HE = John]

A similar observation has been made by Prince (1981) with respect to the example in (13), repeated below as (18):

(18) Paul called Jim a Republican. Then HE insulted HIM.
[Jim insulted Paul]

In (19) we get the same stress pattern as in (18), despite the fact that the two pronouns do not have the same range of possible values (cf. Prince, 1981):

(19) Paul called Jane a Republican. Then SHE insulted HIM.

The stress on the pronouns in (18) as well as (19) evokes contrast, rather than a shift in preferred reference. That is, as was pointed out with respect to (7c) above, when the two pronominal arguments are stressed, the situation described by the argument structure is contrasted with the situation described by the reversed argument order in the preceding clause. Again – like in (15) – when we replace the pronouns by proper names, the stress is preferably maintained: *Paul called Jim a Republican. Then JIM insulted PAUL*. Hence, we get the implicature *And not the other way around* both in (18) and in (19). That explains that *to call someone a Republican* is interpreted as an instantiation of insult in (18) and (19), but not in (12) above with the unstressed counterparts of the two pronouns. This is further illustrated by using either two identical predicates in (20) or two different predicates in (21). In (21), as in (22), the use of stress is not necessary, whereas in (20) it is.

(20) Paul insulted Jane. Then SHE insulted HIM.

(21) Paul called Jane a Republican. Then she HIT him.[5]

(22) When she threatened him with her womanhood, he hated her.[6]

Because there is no contrastive relation between the situation described by the *when*-clause and the situation described by the main clause, the reversed order of the pronouns in the second clause does not have to be marked by stress (*she-him* versus *he-her*) in (22). The contrast evoking function of stress is also obvious in the following examples from Postal (1972) where in each case, only one value for the pronoun is available, and the stress merely signals contrast between the situation described in the sentence and the situation described by the conversational implicature. Without further context, we come up with an implicature like *somebody else is a dope* in (23). In (24)–(27) the contrasted situations are not just implicated, but are part of the meaning via the quantificational (focus association) elements such as *only* and *no other than*. So, we get *the others in our class don't have telepathic powers, others didn't agree to defend that theory, all the others agreed to forget that,* and *other people don't put ketchup on their cornflakes*, respectively:

(23) Melvin, and *HE* is no dope, thinks that the proof is correct.

(24) Melvin, and only *HE* of those in our class, has revealed telepathic powers.

(25) Joan, but no other than *SHE*, has agreed to defend that theory.

(26) Except for Bob, and I am not even sure of *HIM*, we all agreed to forget that.

(27) Tony and *HE* alone, puts ketchup on his cornflakes.

At this point, reconsider Kameyama's example (14), repeated as (28) below:

(28) Jack and Mary are good friends. {He/HE} is from Louisiana.
[He/HE = Jack]

In this example again, the stress signals a contrast between two situations. When the pronoun *he* is stressed, the sentence in (28) gives rise to the conversational implicature that Mary is *not* from Louisiana. In the following example from Bosch (1983), the two situations described by the coordinated main clauses are similar in that they are both anaphoric to the rhetorical antecedent event described by the *when*-clause (cf. de Hoop and de Swart, 2000); the contrast is between the two (shifted) topics involved in the anaphoric events, the male and the female. This contrast is marked by the stressed pronouns:

(29) When the Smiths arrived, HE waited in the car and SHE rang the bell.

4 *Fire from Heaven*

So far, I have pointed out that in all those cases where no alternative values for stressed pronouns are available, the Complementary Preference Hypothesis cannot be applied. Instead, the stress evokes a rhetorical contrast between situations that are around in the discourse or implied by the content of the utterance. With respect to the examples of Kameyama, where alternative values are in fact available, I claimed that the contrastive reading is maintained. But, of course, from this I cannot conclude that the Complementary Preference Hypothesis should be rejected. It could well be that both the Complementary Preference Hypothesis and a contrast evoking condition constrain the interpretation of stressed pronouns in discourse. In order to reject the Complementary Preference Hypothesis, we should look for examples where the Complementary Preference Hypothesis *can* be applied but does not give the right results (in the sense that the predicted value for a stressed pronoun does not correspond with the actual interpretation), or ideally, where we observe a clear conflict between a reading predicted by the Complementary Preference Hypothesis and one predicted by a condition dealing with the rhetorical relation of contrast. In this section I will argue that such examples indeed exist and that the conflict is resolved in favor of the contrastive reading.

In the novel *Fire from Heaven* by Mary Renault I found 50 examples of stressed pronouns, indicated by the author by means of italics (replaced by capitals by me). For the vast majority of these examples, it can be argued that the stress signals contrastive focus. On the other hand, none of these examples can be explained by the Complementary Preference Hypothesis alone. The following example illustrates once more the event in which the Complementary Preference Hypothesis cannot account for the stress as there is no alternative value for the second person pronoun:

(30) 'Well, it teaches you to bear your wounds when you go to war.'
'War? But you're only six.'
'Of course not, I'm eight next Lion Month. You can see that.'
'So am I. But YOU don't look it, you look six.'
(YOU = the addressee)

In (30) the stress signals a contrast between the situations that the addressee doesn't look eight and the speaker himself who does look eight.

Some other examples where the Complementary Preference Hypothesis cannot apply are (31)–(33) with a stressed third person pronoun:

(31) 'Of course', he said. 'I shall kill Attalos as soon as I can do it. It will be best in Asia.' Hephaistion nodded; he himself, at nineteen, had long lost count of men he had already killed. 'Yes, he's your mortal enemy; you'll have to get rid of HIM. The girl's nothing then, the King will find another as soon as he's on campaign.'
(HIM = Attalos)

(32) His mother had risen on one elbow, with the clothes pulled up to her chin. 'No, Philip. Not tonight. It is not the time.' The King took a stride towards the bed. 'Not the time?' he said loudly. He was still

panting from the stairs on a full stomach. 'You said that half a month ago. Do you think I can't count, you Molossian bitch?' The child felt his mother's hand, which had been curved around his body, clench into a fist. When she spoke again it was in her fighting voice. 'Count, you wineskin? You're not fit to know summer from winter. Go to your minion. Any day of the month is the same to HIM.'

(HIM = your minion)

(33) 'I left Oxhead in the road outside. Will you see him safe for me? Take a guard or four.'
'Yes, Alexander.' He went off in a blaze of gratitude.
There was a felt silence; Antipatros was looking oddly under his brows. 'Alexander. The Queen your mother is in the theatre. Had SHE not better have a guard?'

(SHE = the Queen your mother)

In the dialogue in (31) there is only one male individual in the third person figuring in the conversation, namely Attalos. Yet, he is referred to by a stressed pronoun. Here, the contrast evoked by the stressed pronoun is between the situation that the addressee has to get rid of Attalos and the situation also available in the discourse that he has to get rid of the girl. That is, the conversational implicature that the stressed pronoun gives rise to is that the addressee does *not* have to get rid of the girl. In (32) the contrast is between two situations, both accessible in the discourse, the situation of the mother to whom the days of the month are *not* the same and the situation of the minion to whom any day of the month is the same. In (33), the situations between which a contrast is established are again both available in the discourse; namely, the situation when Oxhead (a horse) gets a guard versus the situation when the queen gets a guard.

In (31)–(33) above, again the Complementary Preference Hypothesis cannot be applied as there are no alternative referents available for the unstressed counterparts of the pronoun. In the following text fragment, two instances of stressed pronouns provide direct evidence against the Complementary Preference Hypothesis:

(34) 'So, think which of them can't afford to wait. Alexander can. Philip's seed tends to girls, as everyone knows. Even if Eurydike throws a boy, let the King say what he likes while he lives, but if he dies, the Macedonians won't accept an heir under fighting age; HE should know that. But Olympias, now, that's another matter. SHE can't wait.'

[HE = Philip/the King; SHE = Olympias]

In (34) there are two referents available for the masculine pronoun (namely, Alexander and Philip) and two for the feminine pronoun (namely, Eurydike and Olympias). As far as I can see, the Complementary Preference Hypothesis would predict *HE* to refer to Alexander and *SHE* to Eurydike. Indeed, both Philip and Olympias are straightforward continuing topics at that point in the text. As a consequence, they would count as the preferred values for the unstressed pronouns *he* and *she*. This would automatically leave Alexander and Eurydike as the preferred complementary values for the stressed pronouns. Neither prediction is borne out, however. In other words, the stressed pronouns do not indicate a complementary preference in reference compared to their unstressed counterparts. Clearly, the Complementary Preference Hypothesis is overruled twice. At the same time, the hypothesis that stressed pronouns signal a rhetorical relation of contrast, can still be maintained. Although this is not immediately clear from the direct context, it is known to the reader of the novel that Philip himself became a king when the Macedonians wouldn't accept an under-age heir. So, of all people, HE should know. The contrast is between him knowing and other people maybe not knowing. The stress on *SHE* signals the contrast between Olympias who cannot wait and Alexander who can (the latter situation is available in the preceding discourse). Note that we are dealing with a male and a female referent here, whose circumstances are contrasted (one can wait, the other can't), but who are certainly not in the same set of possible values for the pronoun *SHE*. The only other possible value for the pronoun *she* would be Eurydike, but the stress certainly is not meant to shift the preferred reference from Olympias to Eurydike. Thus, the fragment in (34) provides two clear pieces of evidence against the Complementary Preference Hypothesis. I conclude therefore that the Complementary Preference Hypothesis is neither sufficient nor necessary for a proper analysis of the interpretation of stressed pronouns in discourse.

Additional support for the latter claim is obtained from the work of Venditti, Stone, Nanda and Tepper (2002) who report the results from an experimental study of on-line interpretation of stressed subject pronouns in English. They model the interpretation of stressed pronouns as a side-effect of establishing a coherent structure for discourse, basically following Kehler (2002). Venditti and colleagues find that accent alone is not sufficient to switch reference to a less salient entity. Rather, the type of inferred coherence relation and the ability of the listener to resolve the presupposition of contrast determines interpretation.

5 An optimality theoretic analysis

A theory that tries to derive the interpretation of anaphoric expressions from constraint interaction is Optimality Theoretic Semantics (cf. Hendriks and de Hoop, 2001). In this theory each utterance is associated with an in principle infinite number of interpretations. Hearers arrive – as fast as they do – at one or two optimal interpretations of the utterance by evaluating the candidate interpretations with respect to a set of (conflicting) constraints. The interpretation that arises for an utterance within a certain context maximizes the degree of constraint satisfaction and is, as a consequence, the best alternative (hence, optimal interpretation) among the set of possible interpretations.

The optimal interpretations that are assigned to stressed pronouns in discourse can be analysed in terms of two ranked constraints. The constraints are formulated below:

(35) *Continuing Topic (CT)*: A pronoun is interpreted as a continuing topic.

(36) *Contrastive Stress (CS)*: Stress on a pronoun indicates a rhetorical relation of Contrast.

Continuing Topic can be seen as the interpretive counterpart of *PRO-TOP* that states that the topic is pronominalized (Beaver, to appear). *Contrastive Stress* is also an interpretive constraint in the sense that the direction of optimization goes from form, a stressed pronoun, to interpretation, here Contrast. At least this version of *Contrastive Stress* does not tell us anything about the other direction of optimization. That is, in this chapter I am not concerned with the question when a speaker should use a certain form (for example, a stressed pronoun) to mark a certain interpretation (e.g., Contrast). The rhetorical relation *Contrast* is defined in Mann and Thompson (1988) as a multi-nuclear rhetorical relation with no more than two nuclei such that the situations presented in these two nuclei are: (a) comprehended as the same in many respects; (b) comprehended as differing in a few respects; and (c) compared with respect to one or more of these differenes. According to Mann and Thompson, the effect of Contrast is that the reader recognizes the comparability and the difference(s) yielded by the comparison being made. Asher (1999) investigates the interaction between discourse structure and stress. He discusses some examples that suggest that stressing has the same effect as using a contrastive particle (such as *but*), introducing the rhetorical relation Contrast. Although I formulated *Contrastive Stress* only with respect to pronouns, it will be clear that it should be generalized to narrow or constituent stress in general (cf. Rooth, 1992). Consider the following example:

(37) Setting: Mats, Steve and Paul took a calculus test. After the grading, George asks Mats how it went.

Q: How did it go?
A1: Well, I passed
A2: Well, I [PASSED]$_F$
A3: Well, [I]$_F$ passed

As Rooth (1992) points out, in A1, uttered with a default intonation contour, and no particular prominence on any constituent, Mats's answer provides a neutral description of the situation, with no specific meaning effects related to focus. In A3, as we know by now, the stress on the pronoun indicates a contrast between situations. Hence, we obtain the conversational implicature that the others – Steve and Paul – did not pass. But this effect is not restricted to stressed pronouns. In A2, by stressing *passed*, Mats suggests that he did no better than passing. That is, there is a conversational implicature that Mats did not ace. In this chapter, I deal only with the contrastive effect of stress on pronouns. I hope it goes without saying that when a stressed pronoun establishes a rhetorical relation of Contrast, the two situations differ exactly with respect to the element that gets contrastively focused, hence the referent the stressed pronoun refers to. Therefore, in A3 above, the situation in which I passed is contrasted with the situation in which the others passed (and not with the situation in which I aced).

Let us now see how these two constraints *Continuing Topic* and *Contrastive Stress* account for the right interpretations of stressed pronouns in case of the fragment in (34) above, repeated below as (38):

(38) 'So, think which of them can't afford to wait. Alexander can. Philip's seed tends to girls, as everyone knows. Even if Eurydike throws a boy, let the King say what he likes while he lives, but if he dies, the Macedonians won't accept an heir under fighting age; HE should know that. But Olympias, now, that's another matter. SHE can't wait.'
[HE = Philip/the King; SHE = Olympias]

The optimal interpretations for the stressed pronouns in (38) follow when both constraints *CT* and *CS* are satisfied, as shown in the tableau in (39):

(39) Constraint tableau for the interpretations of stressed pronouns[7]

Input	Output	Contrastive Stress	Continuing Topic
HE in (38) ☞	HE = Philip		
	HE = Alexander		*!
SHE in (38) ☞	SHE = Olympias		
	SHE = Eurydike		*!

The optimal interpretations are the ones we obtain when both *Contrastive Stress* and *Continuing Topic* are satisfied (indicated by the pointing finger in the tableau). In these optimal interpretations, we establish a rhetorical relation of Contrast induced by the stressed pronouns and we interpret the pronouns as continuing topics. Note by the way that in order to satisfy the Complementary Preference Hypothesis and shift reference, we would have to violate *Continuing Topic*. That means that either the Complementary Preference Hypothesis in its present form does not exist at all, or if it does exist, it must be a weak constraint, weaker at least than *Continuing Topic*. For the moment I will proceed from the hypothesis that we can do without the Complementary Preference Hypothesis altogether. This suggests that the combination of *Contrastive Stress* and *Continuing Topic* appropriately constrains the interpretation of the stressed pronouns in discourse. Note, however, that the result with respect to the stressed pronouns in (37) does not give us any clue as to the ranking of the two constraints. The order in the tableau in (39) is not meant to indicate a ranking between them. Obviously, cases in which the optimal candidate does not violate any constraints at all do not allow us to determine the ranking within a given set of constraints. In general, we need to look at cases that involve a conflict between constraints in order to determine the ranking.

A case in point may be examples such as the (13) = (18) and (20) above. In these cases we may argue that *Continuing Topic* is violated.[8] Indeed, the violation seems to be triggered by the necessity of establishing a rhetorical relation of contrast between two similar situations, as required by the stressed constituents. That means that those cases provide evidence for the ranking *Contrastive Stress* >> *Continuing Topic* as illustrated by the tableau:

(40) Constraint tableau for the interpretations of stressed pronouns

Input	Output	Contrastive Stress	Continuing Topic
(18) Paul called Jim a Republican. Then HE insulted HIM.	*HE = Paul;* *HIM = Jim*	*!	*
☞	*HE = Jim;* *HIM = Paul*		**
(20) Paul ☞ *insulted Jane. Then SHE* insulted HIM.	*SHE = Jane;* *HIM = Paul*		**

In fact, we might say that the optimal interpretation of the stressed pronouns in (18) satisfies the Complementary Preference Hypothesis. This is possible because in this example satisfaction of the Complementary Preference Hypothesis corresponds with satisfaction of *Contrastive Stress*, while *Continuing Topic* (a constraint that would be stronger than the Complementary Preference Hypothesis, but weaker than *Contrastive Stress*) is violated by the winning interpretation. Apparently, in a small subset of examples *Contrastive Stress* coincides with complementary preference. However, we do not actually need the Complementary Preference Hypothesis in order to account for the right interpretations of the stressed pronouns in (18). The combination of *Contrastive Stress* and *Continuing Topic* suffices. Moreover, the Complementary Preference Hypothesis would not contribute to computing the right interpretations of the stressed pronouns in (20) as in that sentence the Complementary Preference Hypothesis cannot be applied, although we do get the same type of (contrastive) interpretation in both (18) and (20), a fact that was discussed above.

Let's have one more look at the tableau in (40). If Paul called Jim a Republican and then Paul insulted Jim, the stressed pronouns would not be licensed by *Contrastive Stress*, because there would be no two situations available in the discourse between which a rhetorical relation of contrast could be established. However, if the stressed pronouns refer to Jim and Paul, in that order, *Contrastive Stress* is satisfied. In order to get a contrastive reading between the two situations, they must be sufficiently similar, too. That is why, in the optimal interpretation, to call somebody a Republican must be interpreted as an instantiation of to insult somebody. The contrast is between the two situations: first Paul insulted Jim and then the other way around. The same holds for the optimal interpretation of (20). This is in general the implicature that we get when two pronominal arguments of a predicate are stressed in the absence of further context (as was already observed for the example in (7c) above).

The example in (41), taken from a Dutch newspaper fragment, may serve as a final illustration of this point (again I replaced italics that indicate stress by capitals):

(41) SYDNEY – In de trein die van het Olympic Park terug naar de stad leidt zingen Nederlandse supporters donderdagavond: 'Inge is okay, olé, olé. Inge is okay, olé, olé.' Dat vindt prins Willem-Alexander ook die een uurtje eerder in het Aquatic Centre Inge de Bruijn kussend feliciteert met haar tweede olympische titel.

De Bruijn, tijdens de persconferentie na afloop van haar winnende 100 vrij, over dat ene moment waarop zij in Sydney nu eens níet het initiatief had: 'Nee, nee, HIJ kuste MIJ.'

(*de Volkskrant*, 22-09-2000)

SYDNEY – In the train returning to town from the Olympic Park, Dutch fans sang "Inge is OK, ole ole ole, Inge is OK, ole ole ole…" on

> Thursday night. Prince Willem-Alexander, who congratulated Inge de Bruijn on her second olympic title with a kiss one hour earlier in the Aquatic Centre, thinks so too.
>
> During the press conference following her victory in the 100 meter freestyle, De Bruijn spoke about the one moment in Sydney in which she did NOT take the initiative: “No, no, HE kissed ME.”

Once more, we get the right interpretation via *Contrastive Stress*. The interpretation is that Willem-Alexander kissed Inge de Bruijn and the implicature says: *And not the other way around.*

6 Conclusion

In general, the existence of (morphological) alternatives raises strong interpretive blocking effects (Blutner, 2000). When there are two forms, it is economical to use them for different interpretations. Thus, when there are two lexical forms for the third person singular, one might be optimally interpreted as a continuing topic, the other one as a deictic topic, a shifted topic or a contrastive one. The same would hold for two pronominal forms distinguished by means of stress. In this paper I have argued that in English stress plays only a minor role in the actual process of pronoun resolution (i.e., disambiguation). Rather, meaning effects of stress on pronouns are general pragmatic effects of constituent stress or narrow focus. In particular, stressed pronouns indicate the presence of a rhetorical relation of contrast between two situations within the discourse. The optimal interpretations that are assigned to stressed pronouns in discourse can be analysed in terms of the interaction of only two constraints and their ranking, such that *Contrastive Stress* >> *Continuing Topic*.

In conclusion, if the makers of *The Conversation* had not used a trick to mislead the audience, the final scenes would have been rather predictable. That is, if ‘us’ had been stressed in the earlier recordings, everybody would have immediately understood the utterance *He’d kill US if he got the chance* to mean no less than *Let’s kill HIM!*

Acknowledgments. I am indebted to the Netherlands Organisation of Scientific Research (NWO) for financial support, grants 051-02-070 for the Cognition project “Conflicts in interpretation” and 220-70-003 for the PIONIER project “Case cross-linguistically.” Thanks for valuable feedback to my fellow researchers on these two projects, to the anonymous reviewer of this chapter, the editors of this volume and to the various audiences at different occasions when I presented this chapter.

Notes

1. Thanks to Ellen Prince for drawing my attention to the role of (un)stressed pronouns in this movie. As she told me, she was so confused when she had first seen (or, rather, heard) this movie in the cinema, that she went back another night, to watch it again and find out about the trick with the (un)stressed pronouns. Thanks to Jennifer Spenader for her help in rewriting this opening paragraph to make it understandable.
2. The examples in (3), (4), and (22) are taken from the novel *Fire from Heaven* by Mary Renault.
3. I do not want to exclude the possibility that in the *presence* of further context (it has to be quite a rich context, then) it is possible to establish contrast between two other situations, for example *HE kissed ME* versus *SHE kissed YOU*.
4. The reviewer points out that for him/her, the preference in (10) is not Bill but John, and the reason would be the presence of *then*. Without *then*, it would be Bill, as claimed by Kameyama. This might well be true, but for the sake of a proper discussion, I will stick to Kameyama's (1999) literal examples and not modify them.
5. The reviewer finds that there has to be a stress on *she* as well: *Paul called Jane a Republican. Then SHE HIT him*. Interestingly, the reviewer points out that the additional stress is not necessary if the second clause is subordinated: *Paul insulted Jane, whereupon she HIT him*.
6. The example in (22) is taken from the novel *Fire from Heaven*.
7. A constraint violation is indicated by an asterisk, while the exclamation mark indicates that a violation is fatal. That is, the candidate is then no longer under consideration.
8. Thanks to the reviewer for pointing out to me that because each pronoun is subject to *Continuing Topic*, the winning candidate violates this constraint twice while the loser only violates it once. Because there is a strict domination hierarchy such that *Continuing Topic* is weaker than *Contrastive Stress*, two violations of *Continuing Topic* is still more harmonic than one violation of *Contrastive Stress*.

3
Optimization in Focus Identification

Petra Hendriks

1 Association with focus

This chapter investigates which factors are involved in the identification of the focused expression with which a focus particle associates. If a hearer wishes to interpret a sentence containing a focus particle, one of the things she must do is identify the focused expression. Although we are concerned here with bound focus only (i.e., the focus with which a focus particle associates, in the terminology of Jackendoff, 1972), bound focus and free focus are traditionally considered to be essentially the same phenomenon. The dominant view seems to be that focus (bound or free) is an abstract feature on syntactic phrases which is marked by prosodic prominence.[1] This abstract focus feature has certain effects either in semantics or in pragmatics, depending on the exact theoretical position. A grammaticalized account of focus such as the structured meaning approach (e.g., von Stechow, 1991; Krifka, 1991) puts much of focus into syntax and semantics. Degrammaticalized accounts of focus such as the alternative semantics approach of Rooth (1992) or the approach of von Fintel (1994), on the other hand, remove focus from the grammar and place it in pragmatics. Under a pragmatic approach, focus is assumed to signal the presence in the context of a certain kind of presupposition, to which focus particles might be anaphorically or presuppositionally related.

Although focus is generally assumed to be marked by prosodic prominence, at the same time it is widely acknowledged that prosodic prominence does not clearly identify and delimit the focus (König, 1991). As an illustration, consider (1) and (2). In these two dialogues, the answers (A) are completely identical. Emphatic stress falls on the direct object (as is indicated by small capitals). However, in (1) focus is generally assumed to be on the direct object *a watch*, whereas in (2) it is assumed to be on the verb phrase *bought a watch*. These different focus assignments are the result of the questions that the sentence provides an answer to. So rather than unambiguously marking the focus, emphatic stress appears to be merely one of the factors involved in marking the focus.

(1) Q: What did Mary buy?
A: Mary only bought a WATCH.

(2) Q: What did Mary do?
A: Mary only bought a WATCH.

In these sentences, focus is determined on the basis of the linguistic context. In fact, Wh-interrogatives are often used as a test for determining the focus of a given sentence in context. With regard to this test, the focus of a sentence can be defined as that part of the sentence that corresponds to the Wh-phrase in an interrogative to which it provides an appropriate answer. However, this test does not always give us the right result. Consider the following dialogue:

(3) Q: Who only bought a watch?
A: MARY only bought a watch.

Here, the subject *Mary* provides the answer to the preceding question. Although *Mary* might be considered the focus of the entire sentence, it cannot be interpreted as the focus with which *only* associates. Because focus is not always formed by the new information in the sentence, as (3) illustrates, context is not able to identify focus correctly in all cases.

More importantly, perhaps, the dialogue in (3) shows that there is no direct tie between emphatic stress and focus. Whereas focus may be on the direct object *a watch* or on the VP *bought a watch* in the answer in (3), the element bearing the main stress of the clause is an entirely different constituent; namely, the subject. These cases of so-called second occurrence focus (see, e.g., Partee, 1999) are highly problematic for any theory of focus that assumes focus to be determined mainly by sentence accent, that is, for almost every current theory of bound focus.

Now let us look at syntactic structure. Would it be possible to define focus in terms of syntactic structure, for example, as the material with which the focus particle combines? The generally accepted view is that this is not the case. In (1), the focus particle combines with the VP *bought a watch*, but nevertheless focus is assumed to be on the noun phrase only. Syntactic structure is not able to distinguish between noun phrase focus in (1) and verb phrase focus in (2). But note that if *only* precedes the subject, as in (4), no amount of emphasis on *a watch* will allow us to interpret *a watch* as the focus of *only*:

(4) Only Mary bought a watch.

Apparently, then, syntactic structure restricts the set of possible foci. However, like prosodic prominence and linguistic context, it does not seem to unambiguously identify the focus.

2 Focus as a semantic property of focus particles

If it is impossible to define the focus in focus particle constructions in terms of either prosodic, pragmatic or syntactic properties, how should we define focus then? The view that is adopted in this chapter is that the focus with which a focus particle associates must be understood as a semantic property which is introduced by the focus particle. That is, a focus particle such as *only* semantically requires focal material to be present in the sentence. In this respect, focus particles resemble quantificational determiners. Quantificational determiners partition the sentence into a restrictor and a nuclear scope. Similarly, focus particles partition the sentence into two parts: the focal part and the non-focal part or background. In Hendriks and de Hoop (2001), it is argued that the two argument sets of a quantificational determiner are determined through the interaction of violable constraints. The central hypothesis of this chapter is that the focus of a focus particle is determined in a similar way.

The proposed account of focus identification is neither a completely semantic one nor a completely pragmatic one. Although I agree with Vallduví (1992), Schwarzschild (1997) and Williams (1997) that *only* does not associate with focus via a compositional mechanism, I disagree with them in the assumption that the lexical entry for *only* does not encode a dependency on focus. As is hypothesized here, focus particles semantically require a focus set and a background set, between which they establish a relation. Which constituents contribute to each set, however, is not determined in a purely compositional way. Syntax plays a role, but only as a soft (i.e., violable) constraint that can be overruled by other, stronger, constraints. Other constraints playing a role in the identification of focus might be prosodic or contextual. In general, the cues by which focus is signalled are assumed to take the form of soft constraints, which can be overruled by stronger constraints. If the interpretation of focus somehow involves a set of alternatives to the focused material, which is a rather uncontroversial assumption (see, e.g., Rooth, 1985), we can then define focus as the part of the sentence that gives rise to this set of alternatives. For an illustration of the basic idea, consider the following example:

(5) Mary bought only a WATCH.

Here, both emphatic stress and syntactic structure point at the phrase *a watch* as the focus of *only*. This focused phrase gives rise to a set of alternatives, for example {a watch, a ring, a book}. The remainder of the sentence yields the other set, here the set of things that Mary bought. So focus particles (FPs) can be seen as establishing a relation between two sets, similar to quantificational determiners:

(6) $\mathrm{FP}_{\mathrm{E}}(\mathrm{A})(\mathrm{B})$

A difference between focus particles and quantificational determiners is that the first argument set of a focus particle (i.e., the set of alternatives) is not simply given by the sentence, but rather is construed on the basis of the focal material which is present in the sentence. But note that the first argument set of a determiner is always construed under the influence of context too (Hendriks and de Hoop, 2001). In Section 4, we will return to the relation between focus and quantification. In particular, we will look at the relation between the quantificational and focus-sensitive properties of *only*. It will be shown here that the relation which a quantificational determiner establishes between its two argument sets is quite similar to the relation which the focus particle *only* establishes between its two argument sets.

Returning to the present discussion, the two sets which form the arguments of the focus particle in sentence (5) are given in (7):

(7) A = {a watch, a ring, a book}
B = λx.buy(m,x)

Here, emphatic stress and syntactic structure pick out the same focused phrase. Often, however, not all cues point into the same direction or are able to unambiguously determine the focus. In the answer in (2), for example, focus is generally assumed to be on the verb phrase. The two argument sets of *only* are therefore the following:

(8) A = {buy a watch, play badminton, read a book}
B = λP.P(m)

Although *only* occurs in VP modifier position, emphatic stress falls on the noun phrase object, as in (5). So the assignment of stress is the same in the answer in (2) as in (5). But whereas in the answer in (2) focus is assumed to be on the VP, verb phrase focus does not seem to be possible in (5). An adequate analysis of focus particle constructions will therefore have to explain how the different factors involved in focus identification interact.

In this chapter focus identification is viewed as a process of optimization, as is characteristic of Optimality Theory (Prince and Smolensky, 1993, 1997). In OT, a grammar consists of a set of well-formedness constraints which apply to structural or semantic representations simultaneously. The constraints are potentially conflicting and are ranked in a hierarchy of relative strength. Conflicts between constraints are resolved because higher ranked constraints have total dominance over lower ranked constraints. Before we turn to the constraints that might be involved in focus identification, let us first look at a number of characteristics of the focus particle *only*.

3 *Only* and conservativity

The main assumption of this chapter is that the focus of a focus particle is determined in the same way as the argument sets of a quantificational determiner. Interestingly, *only* has a dual status. On the one hand, it is a focus particle. At the same time, however, *only* has quantificational properties. Since *only* can appear in determiner position, one would expect *only* to display all properties displayed by quantificational determiners in general. For example, *only* is expected to display the property of conservativity:

(9) Conservativity: $\text{Det}_E(A)(B)$ iff $\text{Det}_E(A)(A \cap B)$

As the validity of the following equivalence shows, the determiner *all* is conservative:

(10) All cats purr ↔ All cats are purring cats

In general, all natural language determiners are assumed to be conservative. As Barwise and Cooper (1981) put it, determiners live on their first argument set. In contrast to other determiners, however, *only* in determiner position does not allow for the equivalence relation in (9):

(11) Only cats purr ←/→ Only cats are purring cats

If it is true that only cats are purring cats, then it is not necessarily true that only cats purr. Because *only* does not appear to be conservative, it has been argued that *only* cannot be a determiner in (11).[2] However, as de Mey (1991) points out, although *only* is not conservative at first sight, it does live on one of its argument sets; namely, its second argument set. De Mey therefore distinguishes between conservativity in the traditional sense, which he terms Right-conservativity, and the type of conservativity that is displayed by *only*, which he calls Left-conservativity:

(12) Right-conservativity: $\text{Det}_E(A)(B)$ iff $\text{Det}_E(A)(A \cap B)$

(13) Left-conservativity: $\text{Det}_E(A)(B)$ iff $\text{Det}_E(A \cap B)(B)$

The following equivalence relation shows that *only* has the property of Left-conservativity and lives on its second argument:

(14) Only cats purr ↔ Only purring cats purr

So *only* in determiner position behaves like a determiner in that it lives on one of its argument sets. But whereas other determiners live on their first argument set (i.e., on the set introduced by the N′ in the above examples), *only* lives on its second argument set (i.e., on the set introduced by the VP in the above examples). We can now use the notion of conservativity and the property of living on an argument set to define the domain of quantification of a quantifier: the domain of quantification of a quantifier is the argument set the quantifier lives on.

4 Focus and quantification

Now why would the focus-sensitive quantifier *only* be Left-conservative, whereas all other quantifiers are Right-conservative? In this section, this will be shown to follow from the view that the focus of a focus particle is determined by various interacting constraints.

Standardly, semantic relations such as the argument sets of a determiner are assumed to be based on syntactic structure. The first argument set of a determiner, that is, the domain of quantification, is supplied by its noun and possible modifiers of the noun. The predicate supplies the second argument set. However, stress can also be a factor in determining the two argument sets of a quantificational determiner:

(15) a. Most ships unload AT NIGHT.
b. Most people SLEEP at night.

The preferred reading of (15a) under the assignment of stress as indicated is that most ships that unload, do it at night. So the first argument set is given by the noun and the verb, whereas the second argument set is given by the adverbial phrase in focus. The preferred reading of (15b), on the other hand, is that what most people do at night is sleep. Here, the first argument set is given by the noun and the adverbial phrase, whereas the second argument set is given by the focused verb. In both examples, non-focal material yields the first argument set of the determiner, that is, the domain of quantification or restrictor. Focal material yields the second argument set of the determiner, that is, the scope of quantification or nuclear scope. If the stress patterns are reversed, we still find this effect:

(16) a. Most ships UNLOAD at night.
b. Most people sleep AT NIGHT.

Here, the domains of quantification are given by the set of ships that do something at night and the set of people that sleep, respectively. That is, the non-focal part of the sentence gives us the first argument set of the determiner. The focal part of the sentence, *unload* and *at night*, respectively, gives us the second argument set of the determiner. This generalization corresponds to Partee's (1991) correlation regarding the relation between focus structure and tripartite quantificational structure: background corresponds to restrictive clause and focus to nuclear scope. According to Partee, this correlation has the status of a default strategy, which can be overridden by explicit syntactic rules in the case of quantificational determiners. In particular, the noun and possible modifiers of the noun always supply the domain of quantification, even if one of these elements is stressed. We will return to this issue in more detail in the next section.

For focus particles, the first argument set is determined by the phrase in focus. So here we have a conflict between the demands of focus and the demands of quantification. The first argument set of a quantificational determiner (the domain of quantification) is given by non-focal material and the second argument set (the scope of quantification) by focal material. In contrast, the first argument set of a focus particle is given by focal material and the second argument set by non-focal material. Because *only* is both a quantificational determiner and a focus particle, this conflict has to be resolved somehow.

Resolution of the conflict between the two roles of *only* can be modeled as a process of optimization. Ideally, hearer optimization proceeds from a contextually enriched acoustic input (the acoustic form of the utterance in combination with the linguistic and extra-linguistic context of the utterance, world knowledge, etc.) and yields a complete semantic representation as its output. For the sake of simplicity, however, I will assume that the hearer has already recognized the speech sounds and assigned a global syntactic structure to the input. Thus, the input of an OT tableau is a syntactically structured sentence in which sentence accents are indicated. The output (i.e., each of the candidates in an OT tableau) is also very much simplified in the analysis presented below and merely consists of a characterization of the quantificational and information structure of the sentence. A final simplification concerns the process of optimization. Although speaker information may also play an important role in hearer optimization (as is formalized in bidirectional OT, cf. Blutner, 2000; Zeevat, 2000), in this chapter interpretation will simply be taken to be a process of unidirectional optimization.

The process of hearer optimization is guided by, among others, the following three soft constraints:

(17) SYNTACTIC STRUCTURE (DET)

If there is an N′ that constitutes an NP together with a determiner, use this N′ to restrict the domain of quantification of that determiner and use the rest of the clause to restrict the scope of quantification of that determiner.

(18) SYNTACTIC STRUCTURE (FP)

If there is an XP to which a focus particle is adjoined, use this XP to restrict the focus of that focus particle and use the rest of the clause to restrict the background of that focus particle.

(19) FOCUSING

> If a constituent contributes to the focus of a focus particle, use this constituent to restrict the scope of quantification of that focus particle and use the rest of the clause to restrict the domain of quantification of that focus particle.

The constraint SYNTACTIC STRUCTURE (DET) is adapted from Hendriks and de Hoop (2001).[3] It requires all material in the N′ to end up in the first argument set of a determiner and the rest of the clause to end up in the second argument set. In a similar fashion, the constraint SYNTACTIC STRUCTURE (FP) makes explicit the role of syntactic structure with respect to the argument sets of a focus particle. It requires all material in the XP sister of the focus particle to yield the focus and all material which is not in the c-command domain of the focus particle to end up in the background set of the focus particle. These two constraints thus partition the sentence into two parts (domain/scope of quantification and focus/background, respectively) on the basis of syntactic structure. Note that it is possible for both constraints to apply to *only* in determiner position because the phrase to which *only* is attached is structurally ambiguous between an N′ and an NP with a null determiner. If *only* appears in some other position than a determiner position, as in *Mary only swims*, the constraint SYNTACTIC STRUCTURE (DET) does not apply. The constraint FOCUSING, finally, reflects the general tendency not to express salient material or introduce new material in the domain of quantification.

If it is assumed that input information such as syntactic structure and sentence accent reappears in the output, these constraints can all be viewed as members of the subclass of markedness constraints. They express the fact that semantic output forms that violate these constraints are more marked than semantic output forms that do not. To determine whether these constraints are violated or not, then, only possible output forms have to be considered.

Given the ranking as in (20), the property of Left-conservativity of *only* follows. According to this ranking, FOCUSING is ranked higher than SYNTACTIC STRUCTURE (FP), and SYNTACTIC STRUCTURE (FP) is ranked higher than SYNTACTIC STRUCTURE (DET):

(20) FOCUSING >> SYNTACTIC STRUCTURE (FP) >> SYNTACTIC STRUCTURE (DET)

Consider the following sentence:

(21) Only CATS purr.

To interpret this sentence, the lexical-semantic properties of *only* require that a certain quantificational structure and information structure be assigned to it. Given the input in (22), we have four possibilities regarding the quantificational structure and informational structure. The noun *cats* may contribute to the domain of quantification or to the scope of quantification. Assuming that constituents must either contribute to the domain or to the scope of quantification and that these two sets may not be empty, this exhausts all possibilities with respect to the quantificational structure of the sentence.[4] In addition, the noun *cats* may either restrict the focus or the background. Since the choice for the noun leaves us no options for the verb and the other way around, this gives us four candidate outputs:

(22) Quantificational structure and information structure of (21)

Input: Only [$_{N'/NP}$ CATS] [$_{VP}$ purr]	FOCUSING	SYNTACTIC STRUCTURE (FP)	SYNTACTIC STRUCTURE (DET)
Q-domain: CATS Focus: CATS	*!*		
Q-domain: CATS Focus: purr		*!*	
☞ Q-domain: purr Focus: CATS			**
Q-domain: purr Focus: purr	*!*	**	**

In the first and the fourth candidate of the tableau, the focused constituent (*cats* and *purr*, respectively) restricts the domain of quantification. Hence, these candidates violate the constraint FOCUSING twice. To see this, consider the first candidate. Here, the focused noun *cats* does not restrict the scope of quantification, thus violating FOCUSING. In addition, the backgrounded verb *purr* does not restrict the domain of quantification. A constraint violation is indicated by an asterisk in the cell belonging to the row of the candidate and the column of the constraint. An exclamation mark indicates a fatal violation of a constraint. A violation is fatal if it renders the candidate suboptimal. A crucial characteristic of the constraints in OT is that they are ranked hierarchically and strictly dominate each other. This means that one violation of a stronger constraint is worse than many violations of a weaker constraint.

In the second and fourth candidate, the verb *purr* restricts the focus of *only*, while the sister of *only* (the noun phrase *cats*) restricts the background of this focus particle. This results in two violations of the constraint SYNTACTIC STRUCTURE (FP): one for the verb and one for the noun.

Because they violate one of the two stronger constraints, the first, second and fourth candidate are all suboptimal. This leaves us with only one candidate; namely, the third candidate. This candidate is the optimal candidate, which is indicated by the pointing finger. According to this candidate, the noun *cats* contributes to the focus of the focus particle. Thus, this candidate satisfies Syntactic Structure (FP). This third candidate also satisfies Focusing because *cats* does not contribute to the domain of quantification of *only*. However, in order to be able to satisfy these two constraints, the weaker constraint Syntactic Structure (Det) must be violated. This explains why the noun *cats* does not supply the domain of quantification of the quantifier *only*. So the interaction among the three constraints introduced above yields an explanation for why *only* lives on its second rather than on its first argument set or, in the terminology of de Mey (1991), why *only* is Left-conservative rather than Right-conservative.

The three constraints introduced in this section also yield an explanation for the interpretation of quantificational sentences with focus-insensitive determiners. If the determiner is focus-insensitive, it does not require a partitioning of the sentence into a focal part and a background part. Hence, the constraints Syntactic Structure (FP) and Focusing do not apply. As the tableau in (24) illustrates, the optimal candidate for (23) is a candidate which complies with the syntactic structure of the sentence.

(23) Most cats PURR.

The result is that the noun *cats* restricts the domain of quantification, while the verb phrase *purr* restricts the scope of quantification:

(24) Quantificational structure and information structure of (23)

Input: Most [$_{N'/NP}$ cats] [$_{VP}$ PURR]	Focusing	Syntactic Structure (FP)	Syntactic Structure (Det)
☞ Q-domain: cats			
Q-domain: PURR			*!*

In this section, an optimality theoretic account was presented of the way in which the quantificational structure and information structure of a focus particle construction are determined. At this point, the proposed analysis does not make any reference to sentential stress. However, as was pointed out in the previous sections, sentential stress does play a role in the identification of focus. Therefore, the next section is concerned with the effects of sentential stress on focus identification.

5 Accenting versus deaccenting

Although determiners such as *most* are believed to be focus-insensitive, emphatic stress can affect the interpretation of quantificational sentences involving these determiners. The effects of stress can be modeled by the following constraint:

(25) DEACCENTING

> If a constituent is anaphorically deaccented, it must contribute to the domain of quantification of a quantifier.

Concerning the status of this constraint, the same considerations that hold for the three constraints that were introduced in the previous section also hold for this constraint. If it is assumed that input information such as sentence accent reappears in the output, this constraint can be viewed as a member of the subclass of markedness constraints as well.

The basic idea with respect to deaccenting is that an element can only be anaphorically deaccented if its sister is contrastively accented (cf. Williams, 1997). So contrastively accenting *large* in the noun phrase *the large ships* gives rise to the anaphoric deaccenting of *ships*. Similarly, contrastively accenting *unload* in the verb phrase *unload at night* gives rise to the anaphoric deaccenting of *at night*. Note that being deaccented is not the same as not bearing any accent. An element is deaccented only if it is the sister of a contrastively accented element. If no contrastive accenting occurs, also no deaccenting occurs. Note, furthermore, that a default accent does not give rise to deaccenting. In cases where default accent is indistinguishable from contrastive accent we expect potential ambiguity, which can only be resolved by contextual information.

The constraint DEACCENTING predicts that in quantificational sentences such as (15a) and (16a), repeated below for convenience, the deaccented part of the VP helps to restrict the domain of quantification:

(15) a. Most ships unload AT NIGHT.

(16) a. Most ships UNLOAD at night.

And indeed, this prediction is borne out by the interpretation of these sentences, as was already pointed out in the previous section. These results follow from the interaction between DEACCENTING and SYNTACTIC STRUCTURE (DET). This is shown in the tableau in (26). Here, candidates differ with respect to whether the phrases *ships*, *unload* and *at night* contribute to the domain of quantification or to the scope of quantification:

(26) Quantificational structure of (15a)

Input: Most [$_{N'}$ ships] [$_{VP}$ unload AT NIGHT]	Deaccenting	Syntactic Structure (Det)
Q-domain: ships	*!	
Q-domain: unload		**!
Q-domain: AT NIGHT	*!	**
☞ Q-domain: ships & unload		*
Q-domain: ships & AT NIGHT	*!	*
Q-domain: unload & AT NIGHT		**!*

Because the adverbial phrase *at night* is accented in (15a), the verb *unload* is deaccented. The deaccented phrase *unload* does not contribute to the domain of quantification in the first, third and fifth candidate, so these candidates violate Deaccenting. The constraint Syntactic Structure (Det) prefers the noun *ships* to contribute to the domain of quantification and the constituents in the verb phrase to contribute to the scope of quantification. Therefore, all but the first candidate violate this constraint once or several times. For example, the second candidate violates this constraint twice because *ships* does not contribute to the domain of quantification and *unload* does not contribute to the scope of quantification. Since the fourth candidate violates only the weaker constraint Syntactic Structure (Det) and only violates this constraint once, this is the optimal candidate. The interpretation of (15a) therefore is the interpretation according to which the noun *ships* and the deaccented verb *unload* restrict the domain of quantification. This can be paraphrased as: most ships that unload, do it at night.

In addition to this interpretation, (15a) has another interpretation. This second interpretation arises if the accent on *at night* is interpreted as a default accent. This is possible because default accents are usually on the rightmost element of a constituent in English. If *at night* bears default accent, no other elements are deaccented, so interpretation simply follows syntactic structure. The resulting interpretation is that what most ships do is unload at night. This interpretation surfaces if no contrast can be established in the context with the accented constituent *at night*.

A similar tableau could be drawn for (16a). The interaction between Deaccenting and Syntactic Structure (Det) yields as the optimal interpretation of (16a) the interpretation according to which the noun *ships* and the deaccented phrase *at night* restrict the domain of quantification.

The interpretation thus is that what most ships do at night is unload. No other interpretations are predicted to be possible.

Given these two constraints, it is predicted that even if an item in the N′ is accented, this accented item is interpreted as contributing to the first argument set. There is no tendency to interpret an accented item in the N′ as contributing to the second argument set.

(27) Most LARGE ships unload at night.

The sentence in (27) cannot be interpreted as meaning that most ships that unload at night are large, or that most ships unload at night and are large. In the proposed analysis, this follows from the fact that the constraint DEACCENTING is formulated as a constraint on deaccented rather than accented material. Because DEACCENTING is formulated as in (25), it does not make any claims about the interpretation of accented material. Therefore, all accented material has to conform to the weaker constraint SYNTACTIC STRUCTURE (DET).

(28) Quantificational structure of (27)

Input: Most [$_{N'}$ LARGE ships] [$_{VP}$ unload at night]	DEACCENTING	SYNTACTIC STRUCTURE (DET)
Q-domain: LARGE	*!	*
Q-domain: ships		*!
Q-domain: unload at night	*!	***
☞ Q-domain: LARGE & ships		
Q-domain: LARGE & unload at night	*!	**
Q-domain: ships & unload at night		*!*

In this tableau, candidates differ with respect to whether the phrases *large*, *ships* and *unload at night* contribute to the domain of quantification or to the scope of quantification. Many more candidates are possible if *unload* and *at night* are allowed to contribute to the argument sets of the determiner separately. The only deaccented element in (27) is *ships*, which is deaccented because *large* is accented. Since no element in the VP is accented, the phrases *unload* and *at night* are not deaccented. The constraint DEACCENTING requires *ships* to be interpreted as contributing to the domain of quantification. All candidates in which *ships* does not contribute to the domain of quantification therefore violate this constraint. Because DEACCENTING does not make any claims about material that is not deaccented or about material that is accented, all other constituents in the sentence have to conform to the constraint SYNTACTIC STRUCTURE (DET). So *large* should contribute to the domain of quantification, whereas *unload* and *at night* should contribute to the scope of quantification. The optimal interpretation therefore is that most ships that are large, unload at night.

Interestingly, a similar effect can be observed with *only*, as was already noted by de Hoop (1995). But here the result is exactly the other way around. Consider the following sentence:

(29) Only LARGE ships unload at night.

This sentence means that only large entities are such that they are ships and unload at night. Because *only* is the inverse of *all*, this corresponds to: all ships that unload at night are large. So deaccented material in the XP to which *only* is adjoined is interpreted as contributing to the domain of quantification. This is exactly as predicted by our constraints, as is illustrated by the tableau below. Note that only a few of the candidates are shown here.

(30) Quantificational structure and information structure of (29)

Input: Only [$_{N'/NP}$ LARGE ships] [$_{VP}$ unload at night]	Deaccenting	Focusing	Syntactic Structure (FP)	Syntactic Structure (Det)
Q-domain: LARGE & ships Focus: unload at night			**!*	
Q-domain: unload at night Focus: LARGE & ships	*!			***
Q-domain: ships Focus: LARGE & ships		*!		*
Q-domain: ships Focus: LARGE		*!	*	*
☞ Q-domain: ships & unload at night Focus: LARGE			*	**
Q-domain: ships & unload at night Focus: LARGE & ships		*!		**
etc.				

Both with focus-insensitive determiners and with *only* we find that deaccented material occurring in a position where it should, according to syntactic structure, contribute to the scope of quantification contributes to the domain of quantification instead. If a focus-insensitive determiner is a determiner of the subject NP, deaccented material in the VP contributes to the argument set expressed by the NP. Since the domain of quantification of *only* as a determiner of the subject NP is provided by the VP rather than the NP, the effect is in the opposite direction. Here, deaccented material in the NP contributes to the argument set expressed by the VP. Under the formulation of DEACCENTING as in (25), this pattern is as expected.

Accented material, on the other hand, is predicted not to contribute to the domain of quantification if it occurs in a position where it should, according to syntactic structure, contribute to the scope of quantification, and vice versa. This prediction seems to be borne out by the following data:

(31) a. Only ships unload AT NIGHT.
b. Only ships UNLOAD at night.

If *only* adjoins to the subject NP, the VP generally yields the domain of quantification. If a constituent in this VP is accented, as in (31), this accented element does not seem to be interpreted as contributing to the scope of quantification. That is, (31a) does not seem to have the interpretation that only ships that do something at night, unload. Similarly, (31b) does not seem to have the interpretation that only ships that unload, do it at night. These interpretations follow from the proposed constraints, as is shown by the tableau below:

(32) Quantificational structure and information structure of (31a)

Input: Only [$_{N'/NP}$ ships] [$_{VP}$ unload AT NIGHT]	DEACCENTING	FOCUSING	SYNTACTIC STRUCTURE (FP)	SYNTACTIC STRUCTURE (DET)
Q-domain: ships Focus: unload & AT NIGHT	*!		***	
Q-domain: ships & unload Focus: unload & AT NIGHT		*!	***	*
Q-domain: ships & unload Focus: AT NIGHT			*!*	*

(32) (Continued)

Q-domain: unload Focus: ships & AT NIGHT			*!	**
Q-domain: unload Focus: ships		*!		**
☞ Q-domain: unload & AT NIGHT Focus: ships				***
etc.				

Indeed, the optimal interpretation of (31a) is that only ships are such that they unload at night or, in other words, that all entities that unload at night are ships.

Summarizing, the following predictions of the proposed analysis were shown to be borne out by the interpretation of relevant examples in English:

(33) Predictions of the proposed analysis:

a. Deaccenting within the second argument of a determiner can affect interpretation.
b. Deaccenting within the first argument of a determiner does not affect interpretation.
c. Deaccenting within the first argument of a focus particle can affect interpretation.
d. Deaccenting within the second argument of a focus particle does not affect interpretation.
e. Accenting never affects interpretation, except indirectly through the deaccenting of other constituents.

The examples presented in this section were all examples with the determiner *most* and the focus particle *only*. However, not all determiners are equally sensitive to sentence accent and not all focus particles have quantificational force. The above analysis therefore only provides a very rough sketch of how the interpretation of quantified expressions and focus particle constructions might proceed. Clearly, more research is needed to determine and explain possible differences among determiners and possible differences among focus particles.

Additional support for our analysis might be provided by data discussed in Beaver and Clark (2001). In general, it is assumed that negative polarity items (NPIs) are licensed in the domain of quantification of a universal quantifier, but not in its scope. Interestingly, NPIs may occur in non-focal

VP positions of the VP modifier *only*, which can be analyzed as a universal quantifier. This is illustrated by the examples below (taken from Beaver and Clark, 2002; see also Horn, 1996; and Herburger, 2000, for a discussion of similar data), where the NPIs *bother*, *give a damn* and *lift a finger* occur inside the VP sister of *only*. Small capitals are mine.

(34) a. People only bother with the MILEAGE.
b. I only gave a damn because I thought YOU did.
c. Faeries would only lift a finger to save their best FRIEND.

The possibility of these NPIs to occur inside the VP sister of *only* follows from the proposed analysis. In these examples, *only* is adjoined to VP. According to the constraint SYNTACTIC STRUCTURE (FP), then, this VP is the focus of *only*. The constraint FOCUSING prefers focal material to be interpreted as restricting the scope of quantification. Hence, the VP is preferably interpreted as contributing to the scope of quantification. In the examples in (34), however, a constituent in the VP is accented. Now suppose the result is that the rest of the VP is deaccented. Deaccented material is interpreted as contributing to the domain of quantification, according to the constraint DEACCENTING. Therefore, the deaccented part of the VP in these examples contributes to the domain of quantification of *only*, despite its occurrence in the scope of *only*. Because NPIs are licensed in the domain of quantification of a universal quantifier, this explains why NPIs are licensed here. So these data seem to support our hypothesis that syntactic constraints on quantificational structure are violable and can be overridden by prosodic constraints.

But note that this explanation of the data in (34) rests on the assumption that the NPIs in the VP are deaccented because some other element in the VP is accented. By using deaccenting in this way, however, we seem to have stretched our earlier definition of deaccenting somewhat. Clearly, *their best friend* is not a sister of *lift a finger* in (34c). But in English, usually only the rightmost element of a contrastively accented constituent is marked prosodically. Therefore, it might very well be that not just *their best friend*, but in fact the entire infinitival clause bears contrastive accent. This would then explain why *lift a finger* is deaccented. Although this might yield a satisfactory explanation for the presence of the NPIs in the focus particle constructions in (34), the exact conditions under which accenting and deaccenting can take place certainly require further investigation.

6 Implications of the proposed account

In this chapter it was argued that the concept of optimization, as it features in Optimality Theory, provides us with a fruitful way of looking at issues of interpretation. As was shown in the previous sections, the conflict that arises as a result of the two different roles of *only* (namely, as a quantificational determiner and as a focus particle) can be resolved by viewing the constraints on determiner interpretation and focus identification as violable. Under the assumption that the constraints governing what goes into the two sets of a focus particle are stronger than the constraint that governs what goes into the two sets of a determiner, it is explained why *only* lives on its second argument set rather than on its first argument set. As was mentioned earlier, the requirement of a determiner or focus particle to establish a relation between two argument sets is part of its lexical-semantic specification. Because of this semantic requirement, sentences containing these elements must have a certain quantificational structure or information structure. How this abstract semantic/pragmatic structure exactly looks like it does in the output is the result of the interaction among constraints pertaining to quantificational or information-structural aspects of the sentence. Quantificational structure and information structure thus need not be specified as separate levels of semantic representation. Rather, they are evoked by certain lexical items and compete for their specification in the semantic representation of the sentence.

The proposed account of focus identification results in a different view on the relation between the focus particle and its focus. Many analyses of focus distinguish between the syntactic domain of the focus particle and the focus with which the focus particle associates. The syntactic domain of a focus particle is defined as the phrase which is c-commanded by the focus particle. In the simplest case, the syntactic domain is assumed to coincide with the focus. However, it is also assumed to be possible for the focus to be a proper subpart of the syntactic domain.

(35) a. John would invite only [$_{\text{NP}}$ MARY]
b. John would only [$_{\text{VP}}$ invite [$_{\text{NP}}$ MARY]]

In (35a), *only* is adjoined to an NP which is assumed to be both the syntactic domain and the focus of the focus particle. In (35b), on the other hand, where *only* is adjoined to the VP *invite Mary*, the focus may be on *Mary*, although it need not. Because focus may project to a higher node, *only* could also be taken to associate with the entire VP in (35b). If the syntactic domain does not coincide with its focus, semantic accounts of focus require some mechanism to relate the focus particle to its focus, for example through complex semantic types (e.g., Rooth, 1985) or through LF movement (e.g., Bayer, 1996).

In the proposed account, on the other hand, the syntactic domain of the focus particle and its focus in principle coincide. This is expressed by the constraint SYNTACTIC STRUCTURE (FP). If a focus particle is adjoined to a phrase, this phrase in principle yields the focus. However, through the interaction of SYNTACTIC STRUCTURE (FP) and DEACCENTING, deaccented material in the syntactic domain may be interpreted as contributing to the background rather than to the focus. So optimization over violable constraints provides us with a mechanism which is strong enough to explain how the focus particle associates with its focus even though the focus particle and its focus might not be adjacent in surface structure. Once we view syntactic constraints as being violable, we do not need any order destroying devices such as movement to explain association with focus. We predict that the distance between the focus particle and its focus and the nature of the material intervening between the focus particle and its focus are only restricted by the possibility of the intervening material to be deaccented and not by constraints on LF movement or semantic restrictions. This might explain why there is some disagreement about the possibility of a narrow focus interpretation if the accented phrase occurs inside a syntactic island. In (36), the accented phrase *a watch* occurs inside a complex noun phrase. Many speakers of English find that this sentence has an interpretation according to which Mary did not revise her decision to buy something else, say a book:

(36) Mary only revised her decision to buy a WATCH.

For other speakers of English, however, such a narrow focus interpretation is impossible. Overt movement out of a complex noun phrase is generally disallowed for all speakers of English. These varying judgements with respect to cases like (36) yield a complication for an LF movement account of association with focus. Alternatively, if the acceptability of sentences like (36) were dependent on the possibility of deaccenting, this variation might be due to subtle differences in (implicit or explicit) context.

In many optimality theoretic analyses of semantic and pragmatic phenomena, syntactic constraints appear to be undominated by non-syntactic constraints. In this chapter, it was argued the syntactic constraints SYNTACTIC STRUCTURE (DET) and SYNTACTIC STRUCTURE (FP) must be dominated by the prosodic constraint DEACCENTING. Since prosodic constraints are able to outrank syntactic constraints, interpretation need not proceed in a strictly compositional fashion. Thus the proposed theory of focus identification corroborates the findings of Hendriks and de Hoop (2001), who argue that the interpretation of quantified expressions is not strictly compositional.

A related issue concerns the modularity of the grammar. If most syntactic constraints are undominated by non-syntactic constraints, and if at the same time the prosodic constraint DEACCENTING outranks the syntactic constraints SYNTACTIC STRUCTURE (DET) and SYNTACTIC STRUCTURE (FP), then linguistic constraints cannot be ordered in a strictly modular fashion. Also problematic for this reason is the currently prevailing view in OT that interpretational optimization is a pragmatic mechanism for completing underspecified linguistic meanings. This view implies that syntactic constraints are always stronger, or more important, than other constraints. However, prosody and context appear to be as important as syntax for the interpretation of a sentence. Interestingly, nothing in the architecture of OT prohibits cross-modular constraint interaction. In fact, a strictly modular interaction among constraints would require extra restrictions on the architecture of the grammar, so it seems. The proposed analysis assumes a very simple architecture for the grammar: the generator and the simultaneously applied constraints establish a mapping between an input representation, which is a syntactic-prosodic form, and an optimal output representation, which is a semantic form. No intermediate levels of representation are assumed or required. The constraints on interpretation refer to syntactic, prosodic or lexical-semantic aspects of the output and can hence be said to be syntactic, prosodic or semantic in nature. However, they do not correspond to different levels of representation, nor are they necessarily ordered in a modular fashion. From an empirical perspective, abandoning the modularity hypothesis might lead to interesting results in other areas of semantics and pragmatics as well. However, these questions with respect to compositionality and modularity crucially depend on whether an alternative analysis is possible of the data discussed here in which syntactic constraints are not violated by prosodic ones.

Finally, although the role of linguistic context was not explicitly discussed here, it was pointed out in Section 1 that linguistic context also is an important factor in the identification of focus. Under the proposed account, linguistic context plays an indirect role because it partly determines whether lexical material can be deaccented. A constituent can be deaccented if its neighbour is accented and if it represents 'given' information. When exactly information counts as given is not an easy matter, but see Schwarzschild (1999) for a formalization.

7 Conclusions

In this chapter, an optimality theoretic account was proposed of focus identification. Under the proposed account, focus is understood as a semantic property which is introduced by the focus particle. The focus which the focus particle requires to be present in the output is determined through the interaction among various soft constraints. An important role is played by the prosodic constraint DEACCENTING, which requires anaphorically deaccented constituents to contribute to the domain of quantification of a quantifier. Under the assumption that this prosodic constraint dominates the syntactic constraints which require the argument sets of a determiner or focus particle to be determined strictly compositionally, an explanation can be provided for the interpretation of focus particle constructions and quantified expressions. In particular, it is explained why certain lexical material in the c-command domain of a quantificational focus particle and in the second argument set of a quantificational determiner can be interpreted as contributing to the domain of quantification. Because focus is taken to be only indirectly related to sentence accent, a clear advantage of this approach is that cases of second occurrence focus do not pose any problems. Also, an explanation can be offered for the well-known observation that the focus-sensitive determiner *only* is not conservative in the standard sense.

Acknowledgements. I thank the audience of the Workshop "Optimality Theory and Pragmatics" in Berlin, the Semantics Club in Groningen, the team members of the NWO/Cognition project "Conflicts in Interpretation", in particular Gerlof Bouma, Helen de Hoop, Irene Krämer, Henriëtte de Swart and Joost Zwarts, and two anonymous reviewers for their valuable comments. I gratefully acknowledge the Netherlands Organization for Scientific Research, NWO (grants 051.02.070 and 015.001.103).

Notes

1. For example, Hoeksema and Zwarts (1991, p. 52) define a focused expression as an expression which "has an accentual peak or stress which is used to contrast or compare this item either explicitly or implicitly with a set of alternatives". According to Beaver and Clark (2002, p. 15), focus is "a perceptible pitch rise on a stressed syllable, in English or Dutch". In many other articles, focus is simply indicated by small capitals, which is implicitly or explicitly assimilated with emphatic stress. In this chapter, we will be careful to distinguish focus from sentential stress.
2. As one of the reviewers remarks, another reason for not considering *only* a determiner is that *only* does not have the same syntactic distribution as determiners. *Only* can combine with proper names (*only Mary*), definite descriptions (*only the women*) and numericals (*only three women*), whereas a determiner such as *most* cannot (**most Mary/*most the women/*most three women*). However, the determiner *all* is also able to combine with definite descriptions (*all the women*) and numericals (*all three women*). Hardly anyone would like to conclude on the basis of these facts that *all* is not a determiner.
3. The original formulation of this constraint is: "If there is an N′ that constitutes an NP together with a determiner, use this N′ to restrict the domain of quantification of that determiner" (Hendriks and de Hoop, 2001, p. 22). Under this formulation, however, the constraint is too weak. It would allow for the possibility that focused material or other non-deaccented material in the VP contributes to the domain of quantification too, contrary to the facts. In this chapter, I have chosen to slightly modify the original constraint. However, another (and perhaps preferable) option would have been to add a weaker constraint stating that all constituents must be used to restrict the scope of quantification of the determiner. A similar choice can be made with respect to the syntactic constraint on focus, SYNTACTIC STRUCTURE (FP).
4. This assumption might be formulated as a constraint on interpretation as well: The argument sets of a determiner or a focus particle may not be empty. This constraint remains undominated in the examples under discussion.

4
Optimality Theoretic Pragmatics and Binding Phenomena

Jason Mattausch

1 Introduction

The purpose of this chapter is to show how recent advances in Optimality Theory can contribute to recent advances in the study of the syntax/pragmatics interface. In particular, I wish to show how proposals of Levinson (2000), which aim toward a pragmatic reduction of Chomsky's Binding Conditions (Chomsky, 1980), can be stated somewhat more elegantly and can potentially be improved upon in other ways when recast in the Bidirectional Optimality Theory advocated by Blutner (2000), Jäger (2002) and Zeevat (2000).

The structure of the chapter is as follows. In Section 2, I summarize Levinson's neo-Gricean theory of generalized conversational implicatures, per Levinson (1983), and discuss how Levinson (2000, ch. 4) relates that theory specifically to patterns of intrasentential anaphoric reference in various languages. His main claim is that (at least some of) the Binding Conditions proposed by Chomsky can be seen as consequences of the presence and interaction of three pragmatic heuristics (namely, Levinson's Gricean-inspired I-, Q- and M-principles) and a secondary claim concerns the compatibility of his reductionist picture with various respectable, evidentially supported hypotheses regarding trends in diachronic change across languages, in particular the evolution of certain patterns of anaphoric reference from certain other patterns.

In Section 3, I will follow Blutner's observation that Levinson's tripartite theory of generalized conversational implicatures (GCIs) can be validated, yet simplified via Bidirectional Optimality Theory, an extension of the Optimality Theory (OT) of Prince and Smolensky (1993/2002). Applying Blutner's insights to Levinson's examples shows that indeed both the rich synchronic data accounted for by Levinson as well as the reasonable diachronic picture he advocates can be captured in Bidirectional OT (Bi-OT), and that this can be done while reducing the ontological commitments of the analysis, specifically by allowing for a dualist opposition of hearer/speaker optimality – *à la* Zipf's 'force of diversification' versus 'force of unification' (Zipf, 1949), Horn's Q- and R-principles (Horn, 1984, 1989), and Levinson's own Q- and I-principles – and deriving the effects of the third component of Levinson's program – the M-principle – directly from the mechanics of Blutner's Bi-OT.

2 Levinson's pragmatic reduction of Binding Conditions

2.1 Background

The most significant precursor to Levinson's GCI theory was the work of Grice (1957, 1975, 1978), who develops a theory of 'non-natural' or non-literal meaning based on the idea that a hearer can (and does) infer information above and beyond what is actually encoded in a linguistic utterance and that he does so in accordance with his own beliefs about the intentions, attitudes and desires of the relevant speaker. The idea of 'non-natural meaning' is dependent on the idea that successful, rational communication requires 'cooperative' behavior. As for what qualifies as being cooperative, Grice gives us four conversational maxims which he believes to be at work and which, as a whole, constitute his *Cooperative Principle*:

Quality
Do not say what you believe to be false.
Do not say what you lack evidence for.

Quantity
Do not say less than is required.
Do not say more than is required.

Relation
Be relevant.

Manner
Avoid obscurity.
Avoid ambiguity.
Be brief.
Be orderly.

Grice argues that one symptom of a speaker and hearer's general, mutual awareness of the Cooperative Principle is the appearance of *conversational implicatures*. A conversational implicature occurs when, given some utterance *U*, a hearer defeasibly infers *P*, where *P* is some proposition that, while not linguistically encoded via *U*, is assumed to be deliberately conveyed. The inference, or implicature, is viewed as the result of a hearer's judgement after evaluating the utterance in light of the conversational maxims. An example adapted from Grice (1975):

(1) I saw Mrs Jones kissing a man in the park.

While no compositional analysis of an utterance like (1) would get us to the conclusion that the speaker knows (or at least believes) that the man he is referring to is not *Mr* Jones, it is clearly the sort of conclusion that language users draw all the time. Levinson's theory of generalized conversational implicatures seeks to identify and formalize the justifications for these types of conclusions.

2.2 Levinson's theory of generalized conversational implicatures

Levinson (1983, 1987b, 1989, 1991, 1995, 2000) argues that the latter three of Grice's conversational maxims can be reduced to three pragmatic principles – the I-, Q- and M-principles. Crucially, each principle involves not only a speaker-oriented maxim, but also a hearer-oriented corollary.

The I-principle is, for a speaker, a maxim of (constrained) minimization, while for a hearer it is a maxim of (constrained) maximization.

I-principle

I(S). 'Say as little as necessary', that is, produce the minimal linguistic information sufficient to achieve your communicational ends (bearing the Q-principle in mind).

I(H). Amplify the informational content of a speaker's utterance, by finding the most specific interpretation, up to what you judge to be the speaker's intended point. Specifically:

a. Assume that stereotypical relations obtain between referents and events unless it is inconsistent with what is taken for granted or the speaker has broken the maxim of minimization by using a prolix expression.
b. Assume the existence or actuality of what a sentence is 'about' if that is consistent with what is taken for granted.
c. Avoid interpretations that multiply entities referred to. Specifically, prefer coreferential readings of reduced NPs (e.g., pronouns or NP-gaps).

I-implicatures are the result of the 'amplification' mentioned in *I*(*H*). For a hearer, semantically specific interpretations are assumed so long as they cohere with background information, presumptions about stereotypical situations, and, of course, any information that might be introduced by a subsequent update. Per *I*(*S*), semantically general statements are preferred wherever semantically less general statements are unnecessary:

(2) John pushed Bill. He fell.

I-implicature: John pushed Bill and then, as a result, Bill fell.

(3) a. A blue four-door Mercedes sedan was stolen from the lot.
b. The vehicle was never recovered.

I-implicature: The aforementioned sedan was never recovered.

The I-principle is systematically tempered by the two remaining principles of GCI theory.

The first of these two, the Q-principle, is, for a speaker, a maxim of informational maximization that restricts the minimization permitted by *I*(*S*) and, for a hearer, a maxim of minimization essentially serving to curb the amplification licensed by *I*(*H*):

> *Q-principle*
>
> *Q(S):* Do not provide a statement that is informationally weaker than your knowledge of the world allows, unless providing a stronger statement would contravene the I-principle.
>
> *Q(H):* Take it that the speaker made the strongest statement consistent with what he knows and therefore that:
>
> a. If the speaker used an expression *W* and *W* forms a Horn scale ⟨*S*,*W*⟩, with an informationally stronger expression *S*, then infer *Know*(¬*S*), that is, that the speaker knows the stronger statement *S* is false.
> b. If the speaker asserted *W* and *W* fails to entail an embedded sentence *P*, which a stronger statement *S* would entail and ⟨*S*,*W*⟩ form a Horn scale, then infer ¬*Know*(*P*), that is, that the speaker does not know whether *P* obtains.

Q-implicatures allow a hearer to infer that if an expression *S* was not used, then the meaning of *S* was not intended, so long as the expression *S* stands in a certain relation to the expression that was actually used, call that expression *W*. Specifically, *S* and *W* must form a *Horn scale*:

> *Horn Scale (Horn, 1972)*
>
> A pair of expressions ⟨*S*,*W*⟩ forms a Horn scale only if:
>
> (i) *S* entails *W* (for some arbitrary sentence frame).
> (ii) *S* and *W* are equally lexicalized.
> (iii) *S* and *W* are 'about' the same semantic relation, or form the same semantic field.

Illustrative examples include *scalar* implicatures coerced, per *Q*(*H*)a, by quantifiers in the appropriate type of opposition with one another:

(4) Some of my friends smoke.
Q-implicature: *Not all* of my friends smoke.

A second group of Q-phenomena, *clausal* implicatures, effected by *Q*(*H*)b, often involve epistemic opposition:

(5) John thinks Mary loves him.
Q-implicature: It is *not* the case that John *knows* Mary loves him.

The importance of restricting Q-implicatures to 'Horn scale pairs' can be appreciated by observing that, from (5), we cannot infer that John does not also believe that his father loves him, as the expressions *Mary* and *his father* do not form a Horn scale (since *Mary* presumably does not entail *his father*, contra clause (i) of the Horn scale criteria).

Crucially, Levinson takes it that Q-implicatures will overrule I-implicatures in cases where they conflict. Thus, a speaker will be allowed to minimize his expression as long as he encodes sufficient information, where "sufficient" means that the I-implicatures induced by the reduced expression will 'fill in' for the hearer whatever gaps the speaker leaves in his message *and* no Q-implicature will be triggered (due to the existence of some comparable, alternative, more informative expression that was *not* used) that would induce an inaccurate interpretation.

Finally, whereas the I-principle can reasonably be called a 'speaker-oriented' decree and the Q-principle can be described as 'hearer-oriented', Levinson's M-principle stands for something of a contract between speaker and hearer to the effect that a speaker is not only allowed (per the I-principle) to minimize his expression somewhat, but he is also *expected* to do so, and where he fails to act in his own 'economy' or 'I-principle oriented' interests, the M-principle demands that a hearer defeasibly infer that there must be some motivation for that failure – namely, the speaker's wish to avoid the (I-)implicatures that would normally be effected by the minimal, sufficient expression.

> *M-Principle*
>
> *M(S):* Do not use a prolix, obscure, or marked expression without a reason.
>
> *M(H):* If the speaker used a prolix or marked expression *M*, he did not mean the same as he would have meant had he used the unmarked expression *U* – specifically, he was trying to avoid the stereotypical associations and I-implicatures of *U*.

Like the Q-principle, the M-principle dominates the I-principle and where M-implicatures are induced, they will generally implicate the negation or complement of the usual I-implicature.

The M-principle is meant to represent what Horn (1984, p. 22) calls "*the division of pragmatic labor*" whereby "unmarked forms tend to be used for unmarked situations and marked forms for marked situations" (Horn, 1984, p. 26). In particular, the M-principle provides empirical coverage for cases of *partial blocking* which – compared to instances of *total blocking*, wherein the existence of a specialized lexical form eclipses completely the availability of some non-specialized expression (cf. *fury/*furiosity*) – are cases where a specialized expression rules out some (usually compound, analytic, or productive) expression for a particular (usually 'normal' or 'stereotypical') subrange of interpretations, but not for the entire range.

Examples of partial blocking are often witnessed in syntax and semantics, see, for instance, Atlas and Levinson (1981) or Horn (1984). One classic example from James McCawley (1978):

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

Here, a simple lexical causative like the one in (6a), can describe a run-of-the-mill act of homicide, whereas the productive causative in (6b) – though unacceptable for describing stereotypical murder, manslaughter, and so on – is not an inappropriate expression assuming that the death being described was a true accident or perhaps the result of a lethal, magic curse.

Thus, the M-implicature triggered by (6b) – and generally any M-implicature – is one which coerces an interpretation of non-stereotypicality due to the use of a marked expression despite the availability of an unmarked one.

2.3 Generalized conversational implicatures and the Binding Conditions

2.3.1 Introduction

Levinson proposes that some of the Binding Conditions of Chomsky's Government and Binding Theory (1980, 1981) follow from patterns of preferred interpretation effected by GCIs.

One version of Chomsky's Binding Conditions – which, in the generative grammar framework in which they were originally proposed, are taken to be innate, inviolable principles of universal grammar – can be stated as follows.

Binding Conditions

Condition A. An 'Anaphor' (reflexive or reciprocal) must be bound in its governing category.[1]

Condition B. A (nonreflexive) pronoun must be free in its governing category.

Condition C. R(eferential)-expressions must be free everywhere.

Because 'preferred interpretations' are, in principle, defeasible, they do not typically render some interpretation impossible for some form, in contrast with syntactic stipulations like the Binding Conditions. Thus, if we find a pattern of anaphoric interpretation in some language that does not appear to be at all defeasible, we are justified in believing that the interpretations which constitute that pattern are not merely 'preferred'. For example, see (7), below:

(7) *John$_i$ is pleased with him$_i$.

However, argues Levinson, we are still entitled to suspect that any 'indefeasible preferences' or tenets of grammar in a language that do not noticeably *conflict* with the defeasible Gricean patterns can be suspected of being manifestations of those patterns. In particular, Levinson hypothesizes that what are, in some languages, seemingly indefeasible, syntactic regulations (like the Binding Conditions) are grammaticalized versions of defeasible preferences, which have 'frozen' or 'fossilized' over the evolutionary history of those languages to the point where they are inviolable rules of the language game.[2]

Insofar as this hypothesis pertains to the effects of the Binding Conditions, one type of supportive evidence we can look for are languages in which typical anaphoric paradigms are merely preferred patterns that have not yet grammaticalized.

Levinson argues that there is plenty such evidence, most prominently languages which lack morphosyntactic means of encoding reflexivity altogether and use pronouns reflexively, thus disobeying Condition B systematically and obeying Condition A only vacuously. His so-called B-then-A account is a story of how, assuming only the three principles of GCI theory to be at work, the effects of Conditions A and B can (over very large periods of time) show up as seemingly unbroken rules of a grammar. The effects of Principle C are derived too in Levinson's program, based on assumptions about the markedness of R-expressions and the influence of M- as well as Q-implicatures.

The B-then-A account gets divided into three diachronic stages: In the initial stage, an analogue to Chomsky's Condition B is expressed as a pragmatic, interpretational rule of thumb, the Disjoint Reference Presumption of Farmer and Harnish (1987) (ostensibly derived from the I-principle), which will in turn effect a reluctance to use ordinary pronouns where locally conjoint reference is intended, in the interest of accurate communication. A second, intermediate stage represents the emergence of specialized, emphatic pronouns, which gradually replace regular pronouns in locally bound contexts. A third and final stage is reached by what Levinson once called 'A-first languages' (see Levinson, 2000, pp. 286–327 for discussion), though they are perhaps best described as B-then-C-then-A languages, since the effect of Condition A is viewed as showing up gradually in a grammar only after Condition B- and C-like effects have been in place for a time. In such languages, Chomskyan 'Anaphors' are evidenced by the appearance of necessarily locally bound reflexives that, over time, come to be preferred over pronominals in whatever environments they (the reflexives) are permitted.

2.3.2 A pragmatic reduction of the Binding Conditions: Levinson's B-then-A account

As the name suggests, Levinson's 'B-then-A' or 'B-first' account takes as a starting point a preexisting anaphoric pattern in which something like Chomsky's Condition B – militating against locally bound pronouns – is present. Whether that pattern exists due to a bona fide grammatical principle or is derived from elsewhere is actually left open, though Levinson suggests that such a principle is at least pragmatically motivated, and is likely derivable from the I-principle (see Levinson, 2000, pp. 329–30 for discussion). In particular, he argues, if 'stereotypical actions' are those performed on an individual distinct from the agent then 'stereotypical' transitive clauses will induce I-implicatures of disjoint reference. The pragmatic analogue for Condition B that Levinson assumes to represent the one stabilized tenet of anaphoric reference in 'B-first' languages is the Disjoint Reference Presumption of Farmer and Harnish (1987):

> Disjoint Reference Presumption (DRP): Clausemate arguments are disjoint unless marked otherwise.

Levinson, following Carden and Stewart (1988), proposes three diachronic stages wherein languages gradually develop reflexives due, according to Levinson, to the original influence of the DRP, plus the subsequent influence of GCIs:

> *Stage 1.* Languages which possess no morphological reflexives, but where disjoint interpretations of core arguments are preferred.
>
> *Stage 2.* Languages in which emphatic pronouns may 'prefer' locally conjoint interpretations.
>
> *Stage 3.* Languages with morphological reflexives, which (either partially or totally) replace pronouns in locally bound environments.

Locally bound pronouns in Stage 1 languages will tend to be interpreted as stereotypically disjoint, per the DRP, and, as a consequence, "only ad hoc means such as the use of an emphatic or marked intonation" (Levinson, 2000, p. 374) can be used to M-implicate the reversal of the DRP, that is, locally conjoint reference.

Levinson (2000) cites a considerable number of examples of languages which appear to do without specialized words or morphemes that encode reflexivity – including Australian languages like Guugu Yimithirr, Austronesian languages such as Fijian, as well as quite a few pidgins and creoles, for example, eighteenth-century Haitian Creole, Palenquero, Guadeloupe, KiNubi and others (pp. 338–41). In these cases, reflexivity is typically expressed by a piece of, say, detransitivizing verbal morphology (like Guugu Yimithirr) or stressed, emphasized, or unreduced object pronoun (like Fijian), which "encourages a coreferential reading" (p. 336), but does not guarantee it.

A language may be said to have reached Stage 2 when it has developed a more or less specialized expression which can be counted on to successfully induce non-stereotypical, especially coreferential, readings. Such expressions are not true reflexives, since they are not necessarily interpreted as locally conjoint. Furthermore, pronouns in Stage 2 languages are still used reflexively. Examples include (according to Carden and Stewart (1988)) Martinique Creole, Mauritian Creole and Bislama.

A further example is English itself, though not its modern form. Specifically, evidence from Old English (see Visser, 1963, pp. 420–39 and Mitchell, 1985, pp. 115–89) shows that the opposition between the OE pronoun *hine* and the emphatic *hine selfne* is not comparable to the opposition between the modern cognates *him* and *himself*, since *hine selfne*, though preferably interpreted as reflexive, did not necessarily induce a locally conjoint interpretation:

(8) Old English (Levinson, 2000, fn. 70/ch. 4, citing Mitchell, 1985, p. 115):

Moyses, se the wæs Gode sua weorth thæt he oft with hine selfne spræe.
Moses$_i$ who was so dear to God$_j$ that he$_i$ often spoke with him-emph$_j$

In Stage 2 languages, the anti-locality (i.e., Condition B-type) effects for pronouns and the locality (i.e., Condition A-type) effects for 'proto-reflexives', for example emphatic pronouns, will start to show up (though defeasibly) by virtue of the DRP and the M-principle. Specifically, the use of an emphatic pronoun where an ordinary pronoun could have been used will M-implicate that stereotypical disjoint reference does not obtain. Note that the judgements below reflect the discussion in Levinson (2000, p. 341), citing Visser (1963), who, in turn, cites Sweet (1882).

(9) Old English (Levinson, 2000, p. 341, citing Visser, 1963, p. 433):

a. He$_i$ ofsticode hine$_{?i/j}$
b. He$_i$ ofsticode hine selfne$_{i/?j}$
'He stabbed him(self).'

Because *M*(*H*) directs a hearer to interpret the marked expression *hine selfne* as a speaker's deliberate avoidance of the 'stereotypical associations and I-implicatures of' the available, unmarked expression *hine*, (9b) will be viewed as a deliberate avoidance of (9a), and thus (9b) will get whatever interpretation (9a) normally does *not* get. Per the DRP, (9a) gets a disjoint reading. Thus, (9b) gets a coreferential reading, per *M*(*H*). Of course, both of these preferences, are, as yet, defeasible; compare the ambiguity of (9a) and of (9b) and (8), above.

A language has reached Stage 3 when the aforementioned emphatic expressions can fairly be said to have grammaticalized into legitimate reflexives, that is, bona fide Chomskyan Anaphors that encode referential dependence in a specific domain. (For, presumably, when morphemes like (modern) English *-self* are learned by children, they are learned not merely as emphatic expressions, but rather as expressions with some actual lexical meaning, see English *self-hatred*, German *Selbstmord* ('suicide'), and so on.) Such expressions are, due to their origin, usually "...marked forms. They tend to be longer, more morphologically complex than ordinary pronouns" (Levinson, 2000, p. 331).

Levinson, following Faltz (1985), eventually makes the general claim that "nearly all reflexives ultimately arise from emphatic or stressed pronouns" (Levinson, 2000, p. 350). It appears, though, that pronouns like OE *hine* became more strongly associated with disjoint interpretations at a very gradual pace. Levinson on Old and Middle English:

> the point at which the emphatic became grammaticalized as a reflexive can perhaps be equated with the point at which it lost its inflection, at the transition of Old to Middle English. But...this long preceded the acquisition of indefeasible (or grammaticalized) Condition B-like patterns outlawing the reflexive use of ordinary pronouns, a practice that survived well into Shakespeare's time and beyond.
>
> (Visser, 1963, p. 435; Haiman, 1995)

Note that the judgements below again reflect Levinson (2000, p. 341, and fn. 69/ch. 4); and Visser (1963, p. 439):

(10) Middle English (Levinson, 2000, p. 341, citing Visser, 1963, p. 421):

a. He$_i$ forseoth hie$_{(?)?i/j}$
b. He$_i$ forseoth hie selfe$_{i/*j}$
'He despises him-emph'

According to Levinson's GCI based analysis, the (preferred) disjoint interpretation of (10a) is now due not only to the DRP and the lack of cancellation thereof by any M-implicature, but also to the influence of a Q-implicature; wherever a reflexive expression could have been used, the use of a pronoun will Q-implicate the inapplicability of a coreferential reading. The Horn scale to be considered here is ⟨*hie selfe, hie*⟩ – *hie selfe* being, according to Levinson, the more informative expression[3] – and thus the use of *hie* in (10a) Q-implicates that the stronger form, *hie selfe*, does not apply (otherwise the speaker would have used it, per *Q*(*S*)). Thus, Q-implicatures here will strengthen the usual M-implicatures and this, according to Levinson, will serve to explain the strong tendency toward disjoint interpretation.

Thus, the (albeit defeasible) Condition A- and B-like effects are, for Levinson, viewed as symptoms of the DRP, M-implicatures and Q-implicatures. He notes that "preferred interpretations...tend to become grammaticalized" (Levinson, 2000, p. 268) and it is true that the preferences in pre-modern forms of English appear to have solidified to the point where they are no longer cancelable preferences, but mandatory interpretational restrictions, see again for example: **Bill$_i$ is pleased with him$_i$.*

It remains to say where Condition C-like effects come from. In languages at each stage, Levinson derives the effect of Condition C in non-local contexts from the complicit pressure of the M-principle and the assumption that, compared to pronouns, full lexical NPs are marked or prolix expressions. Consider:

(11) a. John thinks he is fat.
b. John thinks the man/John is fat.

According to *M*(*H*), if the speaker used a prolix or marked expression *M*, he did not mean the same as he would have had he used the unmarked expression *U*. A hearer is thus to infer that a speaker who uttered (11b) does not mean what he could have expressed with (11a), since he is avoiding (11a) at a cost to himself, presumably to avoid exactly those I-implicatures that (11a) would typically effect (especially coreference).

It is not so easy for Levinson's framework to explain how Condition C-type effects show up in locally bound environments. In Stage 3 languages, Levinson derives Condition C-like effects from the Q-principle, since, as before, where a reflexive is available and not used, disjoint reference is Q-implicated. The Horn scale to be considered is now ⟨*himself, John*⟩, *himself* still being the more informative expression, according to Levinson.[4]

However, as Levinson himself notes, "[p]uzzles remain" (2000, fn. 57/ ch. 4), for, in Stage 1 and 2 languages, it is left to say why R-expressions – marked by assumption – do not M-implicate the reversal of the DRP in the same way emphatic pronouns were argued to. In the discussion that follows, I hope to show how letting Bi-OT do the work of GCI theory will not only allow us to capture any fortunate results with less machinery, but also avoid the puzzles which confront Levinson's approach.

3 Recasting Levinson's account in Bidirectional OT

3.1 Motivations

The purpose of the remarks below is to lay out a strategy, following Blutner, whereby we can let a small set of commonsense generative constraints do the work of the I- and Q-principle speaker maxims, while using a second handful of interpretational constraints to mimic the coverage of the hearer corollaries of those two principles, and then show how the effects of the M-principle can be achieved without further stipulation by virtue of the mechanics of Bi-OT. In addition, I will show how some empirical difficulties that face the GCI approach can be alleviated by using the Bi-OT approach instead.

Blutner (2000) has shown how the M-heuristics of Levinson's program can be deduced from the so-called 'weak' version of Bi-OT, which eschews the (somewhat redundant) speaker maxim of avoiding dispreferred expressions without an overruling reason to do so, and views the effect of the hearer corollary as an epiphenomenon resulting from tension (or compromise) between the Q- and I-heuristics. The first goal is to illustrate how this approach can be applied to the examples of anaphoric patterns discussed above.

3.2 Optimality and super-optimality

Recent work including, but not limited to that of van der Does and de Hoop (1998), de Hoop and de Swart (2000) and Hendriks and de Hoop (2001) has applied Prince and Smolensky's OT – a framework originally proposed as a theory of generative phonology, and later syntax – to semantics.

In OT, a certain input gets associated with a multitude of possible outputs. Each possible output is then evaluated with respect to a series of ranked, violable constraints. The various possible outputs are compared to one another on the basis of which constraints they violate, the relative violability (i.e., ranking) of the constraints, and the number of violations committed in order to determine the 'optimal' or 'maximally harmonic' candidate relative to the original input:

> *Relative harmony* (Prince and Smolensky, 1993)
>
> Relative to a constraint hierarchy, *H*, a candidate, *S*, is more harmonic than a candidate, *S′*, (write: $S \succ_H S'$), if *S* 'better-satisfies' *H*, where "better satisfies *H*" means that *S* commits less violations of a constraint *C* than *S′* does, where *C* is the highest ranked constraint in *H* with respect to which *S* and *S′* differ in their performance.

Constraints in OT inevitably conflict, and it follows from the notion of relative harmony that the avoidance of a violation of one constraint may justify the violation of other constraints.

The major distinction between the first approaches to OT semantics and previous applications of OT to phonology, morphology and syntax was that the semantically geared versions were interpretational, not generative, enterprises and thus the pertinent constraints judged candidate meanings with respect to input forms, not candidate forms with respect to input meanings.

In other recent proposals, Blutner (2000), Jäger (2002), Zeevat (2000), and others have all argued that *bidirectional optimization* – that is, a combination of generative and interpretational optimization – is of central importance if we wish to apply OT to natural language interpretation. With a generative dimension added to the 'traditional' OT semantics framework, another sort of optimality – optimality with respect to both evaluation procedures – may be defined and, with this, one may begin to represent the interdependency of the two dimensions, for it is exactly this interdependency that is the major focus of the Grice and Levinson literature, as it is generally seen as the root cause of most conversational implicatures.

Where we write '$\langle F, M\rangle$' to stand for some form/meaning pair, we can write '$\langle F', M\rangle > \langle F, M\rangle$' to mean that, relative to M, F' is more harmonic than F and '$\langle F, M'\rangle > \langle F, M\rangle$' to mean that, relative to F, M' is more harmonic than M. The definition of bidirectional optimality is then straightforward:

Bidirectional optimality (strong version, Blutner, 2000)
A form/meaning pair, $\langle F, M\rangle$ is bidirectionally optimal iff:

q. There is no distinct pair $\langle F', M\rangle$ such that $\langle F', M\rangle > \langle F, M\rangle$
i. There is no distinct pair $\langle F, M'\rangle$ such that $\langle F, M'\rangle > \langle F, M\rangle$

From the definition above, a pair $\langle F, M\rangle$ satisfies Blutner's '*q*-principle' just in case F is an optimal expression given some semantic input M. On the other hand, a pair $\langle F, M\rangle$ satisfies the '*i*-principle' just in case M is an optimal interpretation of F. We can view the *q*- and *i*-principles as being integral parts of the human strategy of natural language comprehension – the *i*-principle being a strategy for determining preferred interpretations and the *q*-principle being a blocking mechanism that, for each form, disqualifies any interpretation that is more harmonic for some alternative form.

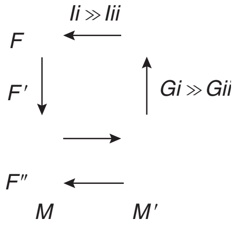
Bi-OT evaluations can be represented in bidirectional *tableaux* which are similar to those of traditional OT but for the presence of a separate tableau for each interpretational possibility. Below, *Ii* and *Iii* represent interpretational constraints and *Gi* and *Gii* are generative constraints. The candidate forms appear, as usual, on the left-hand vertical axis, while the candidate meanings are below, horizontally. Interpretational constraints and generative constraints are assumed not to interact with each other.

Let us assume for the purpose of illustration that $Gi \gg Gii$ and $Ii \gg Iii$:

	Gi	*Gii*	*Ii*	*Iii*	*Gi*	*Gii*	*Ii*	*Iii*
F				*	*		*	*
F′	*		*	*		*	*	
F″	*				*	*	*	
	M				*M′*			

The tableau above represents that $\langle F, M\rangle$ and $\langle F', M'\rangle$ are bidirectionally optimal pairs, whereas F″ is not a member of any bidirectionally optimal pair and thus is disqualified as the output for any (intended) meaning.

The results above (and the results of any Bi-OT analysis) can be represented in 'arrow diagrams', due to Dekker and van Rooy (2000), who note parallels between the Bi-OT literature and work in Game Theory:



Here, the horizontal arrows represent the interpretational preferences relative to the various forms, the arrows pointing to the left showing *M* to be most harmonic for *F* and *F*″, and the arrow pointing to the right signifying that the optimal candidate for *F*′ is *M*′. Likewise, the vertical arrows show the generative preferences relative to the relevant meanings. Here, *F* is the optimal candidate, given *M*, and *F*′ is optimal for *M*′. The absence of any arrow selecting *F*″ means that *F*″ is blocked.

This formulation of bidirectional optimality enables us to model cases of total blocking, where some forms (e.g., **yesterday night, *furiosity*) do not exist because others do (*last night, fury*). However, as was noted above, blocking is not always total, but may be partial. According to the Bi-OT we have considered so far, a pair ⟨*F, M*⟩ is bidirectionally optimal just in case *F* and *M* are optimal *for each other*. However, the fact that *F* is optimal for *M* in such cases is seen as having nothing to do with the fact that *M* is optimal for *F* (and vice versa). In other words, each direction of optimization is independent of the other and the results of optimization under one perspective are not assumed to influence which structures compete under the other perspective.

However, we saw how Levinson's M-principle enabled him to capture cases of partial blocking and the 'marked forms for marked meanings' pattern and – since the primary, initial motivation for developing a bidirectional version of OT was to capture the Gricean and neo-Gricean results heralded in the so-called *radical pragmatics* literature and tradition of Atlas and Levinson (1981), Horn (1984), and others – the situation clearly calls for a version of Bi-OT where the two directions of optimization refer to one another. Such a formalization has been given in Blutner (2000).

Blutner's *weak bidirectional optimality* or *super-optimality* inexorably links the *q*- and *i*-criteria above so that the evaluations that determine optimality for form-for-meaning and meaning-for-form are no longer completely independent of each other, but entirely interdependent:

Bidirectional optimality (weak version)
A form/meaning pair, ⟨*F, M*⟩ is weakly bidirectionally optimal iff:

q. There is no distinct pair ⟨*F*′, *M*⟩ such that ⟨*F*′, *M*⟩ ≻ ⟨*F, M*⟩ and ⟨*F*′, *M*⟩ satisfies *i*.

i. There is no distinct pair $\langle F, M'\rangle$ such that $\langle F, M'\rangle > \langle F, M\rangle$ and $\langle F, M'\rangle$ satisfies q.

The point of the definition above is that for a pair $\langle F,M\rangle$ to *fail* to be super-optimal, it is not enough that there be a distinct pair $\langle F', M\rangle$ or $\langle F, M'\rangle$ that outperforms $\langle F, M\rangle$. Rather, $\langle F, M\rangle$ lacks super-optimal status only if there is a superior pair $\langle F', M\rangle$ or $\langle F, M'\rangle$ *and* the superior pair is itself super-optimal.

At first glance, such a definition might seem a bit bewildering, for the definition for satisfaction of the *q*-condition is included in the definition for satisfaction of *i*-condition, which is in turn included in the definition for satisfaction of the *q*-condition. However, as Jäger, who has explored the formal properties of super-optimal evaluation, points out (Jäger, 2002), the definition is not circular so long as we assume that the '>' relation is a well-founded one.

Consider McCawley's example once again:

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

In order to say why the marked form in (6b) gets associated with a marked meaning, we need to say explicitly why the marked forms and meanings are actually marked. In OT, constraints alone determine what is marked and what is not. We can suppose, then, that (at least) the following two constraints are at work:

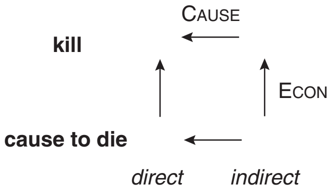
Cause: Interpret causatives directly.

Econ: Avoid productive, compound, or analytic expressions.

The generative constraint Econ will punish the form *cause to die* (under any intended meaning), whereas the interpretational constraint Cause will militate against an indirect-cause reading (given any form). We have:

	Econ	Cause	Econ	Cause
kill				*
cause to die	*		*	*
	direct		*indirect*	

And:



While the tableau and diagram above involve only one strongly bidirectionally optimal pair, both the pair ⟨**kill**, *direct*⟩ and the pair ⟨**cause to die**, *indirect*⟩ are super-optimal. Specifically, though the pair ⟨**cause to die**, *indirect*⟩ is not strongly bidirectionally optimal (since **cause to die** prefers a *direct* interpretation and the *indirect* meaning prefers the more economical form **kill**), that pair is indeed super-optimal, exactly because there is no super-optimal pair ⟨**cause to die**, *M*⟩ such that ⟨**cause to die**, *M*⟩ $>$ ⟨**cause to die**, *indirect*⟩ and there is no super-optimal pair ⟨**F**, *indirect*⟩ such that ⟨**F**, *indirect*⟩ is more harmonic than ⟨**cause to die**, *indirect*⟩.

Below I will show how this version of super-optimality will enable us to derive Levinson's results concerning anaphoric paradigms in natural language.

3.3 Deriving Levinson's results

3.3.1 Introduction

In the remarks that follow, I wish to demonstrate how a small repertoire of constraints, plus the definition of super-optimality given above, can provide us with sufficient means to derive the patterns of local and non-local anaphora captured in Levinson.

3.3.2 Patterns of non-local anaphora

Following Levinson, we can assume that there is some stabilized pattern of anaphoric reference in every language at any given time. It seems that regardless of what pattern a language follows (or what diachronic stage it has reached), it is inevitably the case that anaphoric expressions exist and that they are generally interpreted as having antecedents. We can predict the first part of this pattern – roughly Levinson's *general anaphora pattern* (Levinson, 2000, p. 264), due primarily to the I-principle in his account – by postulating a generative constraint I will call '*FULL':

*FULL: Avoid R-expressions.

The very existence of anaphora in natural language seems to depend on some force or factor such as the one represented by the constraint *FULL and any tenable theory of anaphora will require some analogue to it. If we restrict our attention to single sentence discourses, then the constraint will work very much like Chomsky's Condition C, or Levinson's *I*(*S*). Recall that the effect of Condition C in non-local contexts was 'derivable' in Levinson's program exactly because of the explicit assumption that lexical NPs are dispreferred compared to pronouns. The constraint *Full, and a general lack of rival constraints, will basically reflect the same assumption.

We can predict the second part of the pattern mentioned above (due to *I*(*H*) in Levinson's GCI framework) by postulating an additional, interpretational constraint, adapted from Beaver (to appear):

> Fam-Def: Definites (i.e., names, pronouns, and definite descriptions) have discourse antecedents.

The constraint Fam-Def just represents the (hearer's) tendency to anchor definites and to resolve anaphora. Where there is an occurrence of an anaphoric expression like *him(self)* in a discourse, resolution will always be preferred where it is permitted (by higher ranked agreement constraints and so on). Similar reasoning applies to names or definite descriptions.

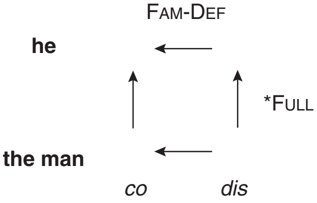
If we consider again the opposition between the following two examples, we can show how the preferred interpretations follow from a Bi-OT analysis involving the two constraints stated above:

(11) a. John thinks he is fat.
     b. John thinks the man is fat.

We have:

	*Full	Fam-Def	*Full	Fam-Def
he				*
the man	*		*	*
	co		*dis*	

The pronoun *he* is the overwhelming favorite here, and, likewise, the conjoint interpretation is the preferred choice given either of the forms. We have:



However, as with the *kill/cause to die* example discussed above, this arrow diagram involves two super-optimal pairs: the strongly bidirectionally optimal ⟨**he**, *co*⟩ and the weakly bidirectionally optimal pair ⟨**the man**, *dis*⟩. Again, the effects of M-implicatures here are being derived without the M-principle, since the pair ⟨**the man**, *dis*⟩ owes its status to the definition of super-optimality and not to an independently stipulated principle of interpretation like *M*(*H*).

In this way we capture Levinson's predictions about the defeasible preference for conjoint interpretations of non-local anaphora as well as the Condition C-like effects illustrated in (11b).

3.3.3 Patterns of local anaphora

Patterns of intrasentential R-expression/pronoun opposition like those exemplified in (11) are much more consistent across languages compared to patterns of reflexive/pronoun opposition. We saw above how Levinson adopts a model of diachronic progression that is divided into three stages which is meant to allow for and explain cross-linguistic differences with respect to patterns of local anaphora while permitting the major players in the analysis – namely, the DRP, the M-principle, and later the Q-principle – to remain the same.

We can capture the effect of the DRP by just assuming some analogue to it to operate as a violable constraint. OT versions of the DRP have been experimented with before, for example, Hendriks and de Hoop's 'Principle B'. However, due to the presence of the "unless ..." clause in Farmer and Harnish's formulation of the DRP, it is more in the spirit of OT to derive the effect of that clause from conflict among multiple constraints. I will use Beaver's constraint 'Disjoint' (Beaver, to appear) as the revised version of the DRP:

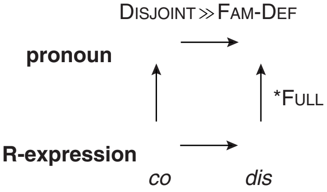
Disjoint: Co-arguments of a predicate are disjoint.

Note that for Disjoint to have any effect on the interpretations of simple transitive clauses, it must be assumed to dominate Fam-Def. (So: Disjoint >> Fam-Def.)

Stage 1 languages. Recall that in Stage 1 languages there are no reflexives. Locally bound pronouns will tend to be interpreted as stereotypically disjoint, per the DRP, but pronouns will nevertheless be used reflexively, *faute de mieux*.

I noted above how, despite the fact that such languages are used as a major piece of evidence for Levinson's GCI-based approach, they cause great difficulty for the approach as well, since there is no obvious explanation for why such languages use pronouns reflexively instead of using full NPs to M-implicate conjoint reference (or, at the very least, why locally bound full NPs do not solicit locally conjoint interpretations). Unfortunately, this problem shows up in the Bi-OT approach as well. For, as the tableau and diagram below indicate, we should expect that, for cases of reflexive transitive clauses, a pair of the form ⟨**R-expression**, *conjoint*⟩ is a superoptimal one:

	*FULL	DIS	FAM-DEF	*FULL	DIS	FAM-DEF
pronoun		*				*
R-expression	*	*		*		*
	co			*dis*		



As far as I can see, there are two possible ways out of this problem.

One option is to simply admit that Chomsky's Condition C (or something like it) is a genuine syntactic principle. This option might seem attractive for, as noted, there are exactly zero human languages that lack anaphoric expressions altogether and we might see this fact as a direct consequence of Condition C being part of the linguistic bioprogram. I refer the reader to Levinson (2000, pp. 298–303), who discusses this hypothesis and summarizes numerous reasons to reject it.

The second possibility is to assume that there really are no such languages that literally lack any means whatsoever of coercing reflexive interpretations, even if those means are not as familiar to us as the '*-self*'-type morphology used in English and even if those means are not totally conventionalized. Virtually all of the languages cited by Levinson evidence the presence of some means – whether a special pronoun, special affix or special syntactic configuration (or perhaps more than one of these) – to induce reflexive interpretations instead of using ordinary, unstressed object pronominals. This assumption is, in effect, a proposal to conflate Stage 1 and Stage 2 languages and take it for granted that every language possesses some method of expressing reflexivity, whether totally ad hoc, totally grammaticalized, or somewhere in between. (E.g., Reinhart and Reuland, 1993, and Reuland, 2001, seem to defend this claim at times.)

Without further argument, I will assume for the time being that the latter possibility is actually the case and leave the defense of that assumption as an area of further research.

With this assumption in hand, we are left with only Stage 2 and Stage 3 languages. I noted above that Stage 2 languages are no less of a problem for

Levinson's GCI-based reductionist strategy than are Stage 1 languages, since there is still no logical explanation of why locally bound R-expressions do not systematically invite coreferential interpretations (via M-implicatures). I will demonstrate below how the Bi-OT approach I have sketched avoids this problem.

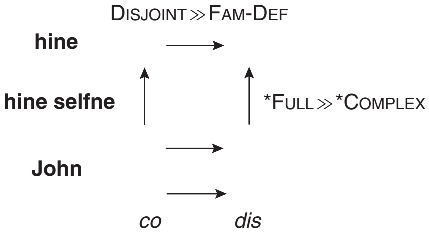
Stage 2 languages. Where available, expressions built up from an emphatic affix combined with a pronoun will compete with ordinary pronouns for the status of ultimate output for the appropriate semantic inputs. Again, Levinson assumes that emphasized, complex, or focused pronouns are dispreferred compared to ordinary pronouns and again I represent this via a markedness constraint on the generative side:

*Complex: NPs are monomorphemic.

The fact that no natural language (to my knowledge) lacks monomorphemic object pronouns (though some do lack object NP-gaps and/or morphologically complex anaphoric expressions) provides evidence for the idea that a constraint like *Complex is a linguistic universal.

Recall example (9), from Old English, and suppose the constraint rankings for OE to be *Full >> *Complex and Disjoint >> Fam-Def. We have:

	*Full	*Complex	Dis	Fam-Def	*Full	*Complex	Dis	Fam-Def
hine			*					*
hine selfne		*	*			*		*
John	*		*		*			*
	co				*dis*			



As before, we no longer need an M-implicature to pair the dispreferred, emphatic pronoun with the dispreferred, conjoint interpretation, since, by definition, ⟨**hine selfne**, *co*⟩ already meets the criteria for a super-optimal solution.

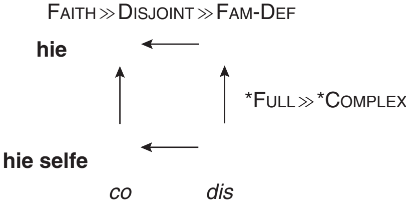
In addition, we now have an answer to the question of why locally bound R-expressions are unacceptable for soliciting locally conjoint readings. In particular, per the diagrams above, the pair ⟨**John**, *co*⟩ does *not* enjoy super-optimal status, since the pair consisting of an emphatic pronoun and the conjoint reading (i.e., ⟨**hine selfne**, *co*⟩, above) is a super-optimal pair and that pair outperforms ⟨**John**, *co*⟩, that is, ⟨**hine selfne**, *co*⟩ > ⟨ **John**, *co*⟩. In this way, relying on Bi-OT allows us to avoid the puzzle that faces Levinson's GCI approach.

Stage 3 languages. The critical issue at Stage 3 is the appearance of morphological reflexives. The distinction between an emphatic like OE *hine selfne* and a grammaticalized reflexive like ME *hie selfe* (or modern English *himself*) being crucial, since the use of the latter will never violate the DRP, due to the "unless…" clause. Because the constraint DISJOINT does not include the "unless…" clause of the DRP, we must derive the effect of that caveat through the interaction of DISJOINT and some other constraint. Let us assume that grammaticalized reflexives are interpreted reflexively by virtue of a faithfulness constraint like the one below:

FAITH: Interpret refexives as locally conjoint.

Assuming the generative constraint rankings for ME to be identical to those of OE and assuming that the interpretational constraint ranking is FAITH ≫ DISJOINT ≫ FAM-DEF, we have the (abbreviated) results below:

	*COMPLEX	FAITH	DIS	FAM-DEF	*COMPLEX	FAITH	DIS	FAM-DEF
hie			*					*
hie selfe	*		*		*	*		*
	co				*dis*			



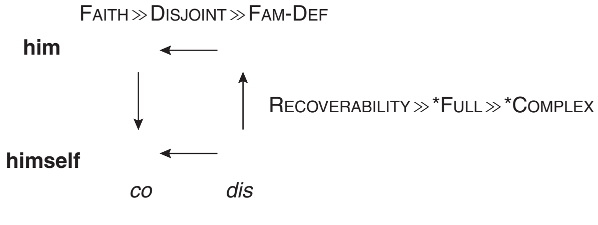
Here, ⟨**hie**, *dis*⟩ is, as before, a strongly bidirectionally optimal pair. And, again, ⟨**hie selfe**, *co*⟩ is weakly bidirectionally optimal. The main difference between languages which lack grammaticalized reflexives and those which have them is that the reflexives are optimally interpreted as locally conjoint, per Faith. Given this state of affairs, a speaker would be ill-advised to use a pronoun where he intended locally conjoint reference or to use a reflexive where he intended disjoint reference, since, in either case, he would almost certainly be misunderstood.

If we follow, for example, Vogel (Chapter 9), and suppose that *recoverability* should be viewed as a necessary condition for grammaticality, then we can model cases like ME or modern English as involving two strongly bidirectionally optimal pairs by adding a constraint to reflect that assumption:

> Recoverability: A form must be optimally interpretable as the meaning it is intended to express.

Assuming that Recoverability is the highest ranking generative constraint, we have:

	Rec	*Complex	Faith	Dis	Fam-Def	Rec	*Complex	Faith	Dis	Fam-Def
him	*			*						*
himself		*		*		*	*	*		*
	co					*dis*				



The results above describe a language in which morphological reflexives are preferred wherever locally conjoint reference is intended, and pronouns are preferred in any case where non-locally conjoint reference is intended. The notion of “non-locally conjoint” is, of course, extremely imprecise since it makes no distinction between conjointness in one sentence, conjointness in a discourse of infinitely many sentences, or two separate discourses. However, if we were to follow, say, Ariel (Ariel, 1990) and assume that some distinction(s) could be made, then we would still have a pattern in which locally bound pronouns, locally free Anaphors, and bound R-expressions are never optimal, and thus will have captured exactly the empirical predictions of Chomsky’s Binding Conditions.

4 'SE-reflexives' and 'long-distance Anaphors'

Levinson has argued that his GCI-based approach to anaphoric paradigms can be extended to offer a great deal more empirical coverage than just to those data discussed above, especially interesting are cases of so-called 'long distance Anaphors' (LDAs) such as those found in Chinese and Icelandic.

(12) Icelandic (Hyams and Sigurjónsdottir, 1990):
Jón$_i$ segir ath María$_j$ elski sig$_{i/j}$.
'Jón says that María loves him/herself.'

The linchpin of his analysis is the assumption that all LDAs invariably involve semantic differences compared to ordinary pronouns, in particular, by soliciting interpretations involving logophoric or "marked deictic perspective" (Levinson, 2000, p. 347).

While I lack the space to provide many details, I have no doubt that if we take Levinson's claim about the semantic distinctions between LDAs and pronouns to heart, extending the Bi-OT analysis above so that it too may account for such cases is unproblematic. For, just as we were able to show why marked emphatic pronouns or reflexive expressions pair with marked locally conjoint readings, we will be able to couple the same marked expressions with (presumably non-stereotypical) logophoric interpretations.

In fact, I believe that a constraint-based analysis like the one outlined in this chapter is probably even better suited to handle such cases, in particular because it can avoid potential controversy surrounding the nature of expressions like Latin/French/Spanish *se* or German *sich*, Norwegian *seg*, Icelandic *sig*, and so on. LDAs are almost inevitably expressions of this sort, that is, monomorphemic expressions which lack number, gender and perhaps person features. (See the potentially long-distance Norwegian *seg* or Icelandic *sig*.)

(13) Icelandic:
Jón$_i$ elskar sig$_i$.
'Jón loves himself.'

Levinson assumes throughout that such expressions are marked for reflexivity and are thus bona fide Chomskyan 'Anaphors' that are 'necessarily referentially dependent' exactly because of that status. This idea has been challenged before (see Bouchard, 1984; Reinhart and Reuland, 1991, 1993; Pollard and Sag, 1992). Reinhart and Reuland (1993), for example, draw the distinction between *SELF* anaphora and *SE* anaphora, the latter being presumed to lack any explicit morphosyntactic feature that indicates reflexivity. It seems reasonable to believe that such expressions might be more appropriately analyzed as *underspecifying* their way to reflexive interpretations as opposed to specifying for them and thus are referentially dependent *by virtue of* their lack of φ-features, rather than the lack of φ-features being a "reflex" of their referential dependence, as Levinson claims (2000, p. 312).

If these challenges are warranted then they would upset the balance of Levinson's GCI-based analysis in a huge way, for if *SE*-type expressions can no longer be considered reflexives or 'heavy' NPs, then it is no longer obvious why they can be said to be 'more marked' or 'more informative' than pronouns and hence the relevant M-implicatures, Horn scales, and Q-implicatures will all be reversed (and, accordingly, pronouns will get reflexive interpretations and the lesser marked, 'lighter' *SE*-type NPs will be assigned the stereotypically disjoint readings).

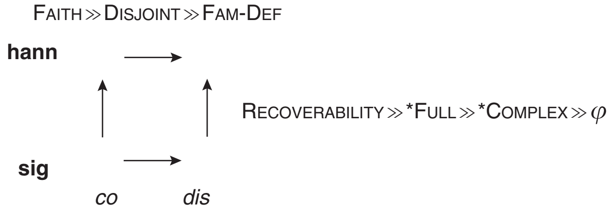
A Bi-OT analysis can avoid this problem completely, for in an OT-based analysis, the notion of markedness is not tied to notions of informativity or 'heaviness' as in Levinson's program. Rather, the definition of markedness is given to us by the constraints alone. If we operate under the assumption that there is some constraint present in all languages that militates against φ-featureless pronouns – and this seems absolutely reasonable, since not a single language on Earth *lacks* φ-feature endowed pronouns – then we have a reason for calling anaphoric expressions which lack such features 'marked':

φ: NPs must have φ-features.

We now have a reason for saying why *SE*-type anaphora take marked meanings, such as reflexive or logophoric interpretations, and we obtain that reason without indulging in the controversial assumption that, say, French *se* or Icelandic *sig* are 'marked for reflexivity'.

In the case of simple, transitive clauses, *SE*-type expressions would come to receive locally coreferential interpretations for the same reason that stressed or emphatic pronouns came to get them, except this time it will be the constraint φ and not the constraint *COMPLEX that is relevant.

	φ	DIS	FAM-DEF	φ	DIS	FAM-DEF
hann		*				*
sig	*	*		*		*
	co			*dis*		


FAITH ≫ DISJOINT ≫ FAM-DEF
hann
sig
RECOVERABILITY ≫ *FULL ≫ *COMPLEX ≫ φ
co
dis

Here, we see that both ⟨**hann**, *dis*⟩ and ⟨**sig**, *co*⟩ are super-optimal pairs.

As noted, Levinson believes that LDAs such as those in Chinese and Icelandic always involve semantic differences compared to ordinary pronouns. For this reason, claims Levinson, the problems that LDAs – especially LDAs like Icelandic *sig*, which exhibits systematic distributional overlap with pronouns (see (12) below) – cause syntactically based theories of anaphora will not trouble a GCI-based theory, since an "Anaphor always contrasts in meaning with the ordinary pronoun, the associated meanings having something to do with emphatic contrast, empathy, or protagonist's perspective, subjective point of view, and so on" (Levinson, 2000, p. 312). Thus, while what Levinson calls an Anaphor – like Icelandic *sig* – may potentially have the same reference as a pronoun, they always contrast semantically on some other level since the 'Anaphor' will carry perspectival information and the pronoun will not. In particular, the logophoric expressions are generally used to indicate that the proposition expressed by a subordinate clause is being reported from the perspective of the subject matrix clause and not the speaker himself. Anaphors, says Levinson (following O'Connor, 1983; Stirling, 1993; and others) are "always referentially dependent and always logophoric". As before, Levinson proposes that the process of becoming 'marked for logophoricity' is – just like the process of becoming marked for reflexivity – a gradual diachronic phenomenon. A Stage 1 language will exhibit the occasional ad hoc intonational or emphatic marking which may be used to – in Levinson's language –'M-implicate' a "marked deictic perspective" (Levinson, 2000, p. 347) or "marked point of view" (p. 348) in contrast with the "unmarked deictic perspective" (p. 348).

If we wish to eliminate the M-principle and derive its effects via Bi-OT, then it is again left for us to state, by means of a constraint, exactly what is "marked" about the perspectival information conveyed by a logophoric expression. Levinson clearly believes that the deictic perspective induced by such an expression is non-stereotypical and there seems little reason to disagree with that intuition. Certainly, the default (read: unmarked) perspective (across languages) is the speaker's perspective. Therefore, we can represent the anti-stereotypicality of 'logophoric perspective' via a constraint which penalizes such interpretations and that constraint will function in a fashion very similar to one we have already seen; namely, DISJOINT:

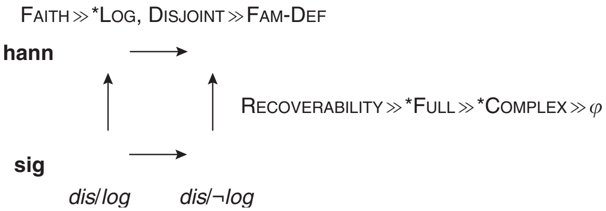
*LOG: Avoid logophoric interpretations.

Consider (12) once again:

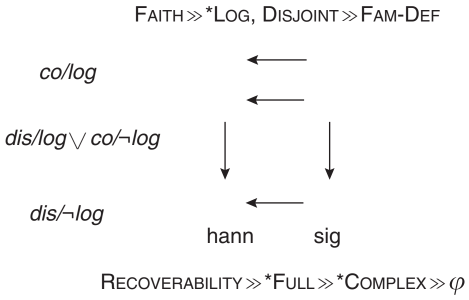
(12) Jón$_i$ segir ath María$_j$ elski sig$_{i/j}$.
'Jón says that María loves him/herself.'

I assume the constraint rankings for Icelandic to be Recoverability ≫ *Full ≫ *Complex ≫ φ and Faith ≫ *Log, Disjoint ≫ Fam-Def. We get:

	φ	*Log	Dis	Fam-Def	φ	*Log	Dis	Fam-Def
hann								
sig	*	*			*			
	dis/log				*dis/¬log*			


Faith ≫ *Log, Disjoint ≫ Fam-Def
hann
Recoverability ≫ *Full ≫ *Complex ≫ φ
sig
dis/log
dis/¬log

Here, the locally disjoint (albeit non-locally conjoint), non-logophoric interpretation and the pronoun *hann* form a strongly bidirectionally optimal pair whereas the logophoric, locally disjoint interpretation forms a super-optimal pair with the expression *sig*. However, as was already pointed out, the interpretation '*dis/log*' is not the only super-optimal partner that *sig* has. For, as was shown above, the pair ⟨**sig**, *co/¬log*⟩ is super-optimal as well. The pairs ⟨**sig**, *co/¬log*⟩ and ⟨**sig**, *dis/log*⟩ have the same relative harmony (due, in particular, to the comparable ranking of the constraints *Log and Disjoint). Thus, a hearer will be inclined to use the expression *sig* if he wishes to induce a locally conjoint interpretation *or* if he wishes to represent a locally disjoint interpretation that involves perspectival information about the subject of some matrix clause:


Faith ≫ *Log, Disjoint ≫ Fam-Def
co/log
dis/log ∨ co/¬log
dis/¬log
hann
sig
Recoverability ≫ *Full ≫ *Complex ≫ φ

Finally, consider once again the case of a simple transitive clause:

(13) Icelandic:
Jón$_i$ elskar sig$_i$.
'Jón loves himself.'

Note that the pair ⟨**sig**, *co/¬log*⟩ will always outperform the pair ⟨**sig**, *co/log*⟩:

	φ	*Log	Dis	Fam-Def	φ	*Log	Dis	Fam-Def
hann		*	*				*	
sig	*	*	*		*		*	
	co/log				*co/¬log*			

This reflects exactly the claim of Levinson and others that where locally bound pronouns contrast with locally bound *SE*-expressions, the semantic contrast will always be a contrast of reference (conjoint versus disjoint) and never of logophoricity (see Levinson, 2000, pp. 323–30 for discussion). The ambiguity of *sig* as deserving a non-stereotypical conjoint interpretation *or* an interpretation involving non-stereotypical perspectival shift will appear *only* in non-locally bound environments.

5 Conclusion

I have tried to show how Levinson's GCI-based approach to a pragmatic reduction of Binding Conditions can be recast in Bidirectional Optimality Theory. As Blutner has already demonstrated, his weak version of Bi-OT is methodologically more austere than GCI theory since the former allows for a bipartite model of neo-Gricean pragmatic principles as opposed to a tripartite one. In addition, aside from harvesting many of the same, fortunate results that Levinson's program is able to capture, the Bi-OT based strategy also overcomes empirical challenges that GCI theory is unequipped to deal with.

Acknowledgements. I am grateful to the organizers and audiences of the 2001 ZAS Semantikzirkel in Berlin, the 2002 Workshop on the Roots of Pragmasemantics in Szklarska Poreba (Poland), and the 2002 Workshop "Optimality Theory and Pragmatics" at the ZAS. Special thanks to Reinhard Blutner, Henk Zeevat and two anonymous referees, who made some very helpful suggestions on earlier versions. All errors are of course my own.

Notes

1. I follow Chomsky and others, and take the ‘*governing category* of X′ to mean the minimal IP- or NP-domain containing X, *bound* to mean coindexed by a c-commanding NP, and *free* to mean not bound.
2. I use the phrase “language game” for the purpose of being deliberately vague, as it is not entirely clear to me what Levinson believes a ‘grammaticalized’ rule of grammar actually is. Since such rules are certainly not innate principles, they must be learned behavior. Thus, it might very well be the case that he believes that the only difference between an non-grammaticalized ‘rule’ and a grammaticalized one is the property of violability or defeasibility which the former possesses and the latter lacks.
3. Levinson, following Popper (1959) and Bar-Hillel and Carnap (1964), argues that an assumption of core-argument coreference increases the informativity of a statement exactly because it restricts the number of entities introduced into the discourse, see Levinson (2000, pp. 273–5).
4. See my Note 3, above.

5
Particles: Presupposition Triggers, Context Markers or Speech Act Markers

Henk Zeevat

1 Introduction

This chapter discusses two possible formal approaches to the semantic/pragmatic characterization of a subclass of modal particles. It may well be that the approaches can be applied to other particles or that they can be applied to certain intonational patterns (e.g., contrastive stress), to morphemes (past tense, agreement) or to words (pronouns) and constructions (some uses of definite descriptions, clefts), but I will not try to to show that here.

The first approach is based on Blutner and Jäger's (1999) optimality theoretic reconstruction of the theory of presupposition that has become fairly standard, and the Heim (1983) and van der Sandt (1992) view of presuppositions as anaphora (see Zeevat, 1992, for an introduction and comparison). The first half of the chapter critically reviews my earlier views on the treatment of particles in this setting, the second part introduces a novel view, again based on optimality theory, which takes as a starting point the marking constraints that are a necessary ingredient of my earlier treatment.

The advantage of the second treatment is not so much that it gives a better account of the particles in question, but that it generalizes better to other particles and that it is more economical. There are more particles that can be seen as context markers than as (non-standard) presupposition triggers.

The empirical content of this chapter is limited to some well-known observations on the English particle *too* in Kripke (ms.), on the Dutch/German particle *toch/doch* (see Karagjosova, 2001) and related particles as in Zeevat (2002).

Discourse particles present a special problem for frameworks in which semantic characterizations are made exclusively in terms of truth conditions. Stalnaker (1973) observed that particles like *even, too, also, doch,* and so on can make the utterances in which they appear pragmatically incorrect, though they can never make a true utterance false. If this is so, it is impossible to deal with the semantic/pragmatic role of particles in such frameworks. Dynamic semantics of the kind that has been assumed for the treatment of anaphora and presupposition is more promising and in fact many particles have been described as presupposition triggers. In a dynamic semantics, meaning becomes a function from an old information state (what the speakers knew already) to a new information state (the old information state together with the information conveyed by the utterance). The characterization of the semantic or pragmatic contribution of discourse particles to the utterances in which they occur is not just a puzzle in pragmatics, but it is also a question with repercussions for the foundations of natural language semantics and pragmatics.

I will conclude that not even the dynamic notion of meaning is sufficient for particles and that a proper account of particles requires more, probably an analysis of speech acts in terms of the conditions under which they can be carried out, and the effects that are achieved if the act is taken seriously by the hearer together with the effects that the speaker intends to achieve. Discourse particles are means for indicating that these are not the normal ones and that other conditions or intended effects apply. The change that a speech act can effect on an information state is only one aspect. In the conclusion, I will give an outline of a theory along these lines.

2 A presupposition theory of certain particles

The particle *too* has occupied a central place in the presupposition literature, both before and after Kripke's underground paper on this particle. The view of Karttunen (1974) is that a presupposition must be true in the context of an utterance of a sentence containing a presupposition trigger that triggers the presupposition if it is not filtered away or stopped by a plug (filters are operators that let through some but not all of the presuppositions of their arguments, plugs are operators that let none of them through). This condition is always met by the simple context of the trigger, like the one in (1).

(1) John will have dinner in New York too.

What is the presupposition? There are a number of readings, but if John carries contrastive stress, it is the statement that somebody different from John will have dinner in New York. Now, New York has many inhabitants and most of them have dinner there every night. In addition, everybody knows that. So in a normal context of utterance, the theory of Karttunen (1974) – and similar theories like Gazdar (1979), Heim (1983), Stalnaker (1973) and van der Sandt (1992) run into the same problem – predicts that the particle *too* cannot change the felicity of the utterance, because its presupposition is trivially met. But the presupposition does matter.

The sentence is infelicitous if the previous conversation has not mentioned another person who will have dinner in New York.

One can try to escape from Kripke's argument by assuming a different presupposition, for example *x* is a person different from John who will have dinner in New York. This is an open formula and can only be satisfied by finding a binder for the *x* in the context: it is very much like a pronoun. Taking *x* as a hidden pronoun has been proposed by van der Sandt and Geurts (2001) in the context of a discourse representation theory. A problem is then that presupposition triggers in theories like van der Sandt's and Heim's generally allow the possibility of accommodation, and the most natural way for applying accommodation leads to regaining the original problematic presupposition: there is somebody apart from John who will have dinner in New York. Van der Sandt and Geurts remedy this problem by assuming that pronouns do not accommodate, something which they motivate by observing that pronouns do not have sufficient descriptive content for accommodation. The observation that pronouns do not accommodate is correct, but the explanation seems problematic, since a pronoun like "he" or "she" has roughly the same descriptive content as "the man" or "the woman" which would accommodate, at least under the assumptions of the theory adopted by van der Sandt and Geurts.

This, however, still allows for partial accommodation: resolve the pronoun to some known entity and accommodate the information that the person will have dinner in New York, as pointed out by Nick Asher.[1] For example (2)

(2) A man is walking in the park. John will have dinner in New York too.

could (must, under the assumptions of van der Sandt and Geurts, 2001) be treated by resolving the pronoun from the presupposition triggered by *too* to the walking man in the first sentence and by accommodating the remaining part of the presupposition, so that it would be equivalent to (3):

(3) A man is walking in the park. He will have dinner in New York. John will have dinner in New York too.

The correctness of Asher's argument follows from a parallel case with an overt pronoun inside the trigger (4), where we indeed seem to accommodate unproblematically the presupposition (the man has a dog) after resolving the pronoun:

(4) A man is walking in the park. Some children are playing with his dog.

The assumption that *too* does not allow accommodation because of a hidden pronoun has another problem as well. The particle *indeed* (or the Dutch *immers*, roughly "As you know") presupposes the content of the whole sentence in which it occurs and so its presupposition has arbitrary amounts of descriptive content. But the presuppositions of these particles cannot be accommodated anymore than the presupposition of *too* and it seems rather artificial to assume a hidden pronoun in the presupposition of *indeed*.

In fact, it is a general property of the particles that are presupposition triggers that their presupposition cannot be accommodated.[2] *Again, indeed, instead,* German/Dutch *doch/toch* and Dutch *immers* are rather clear examples.

The particles also have other properties that make them unlike normal presupposition triggers. First of all, they are not optional in the sense that if one finds them in a body of natural text or dialogue they can just as well be omitted; (5) is an example, but one really needs to look at the total picture:[3]

(5) A: Bill will come tonight.
B: John will come *(too).
A: Bill is ill.
B: He is *(indeed).

Second, they have a rather minimal meaning apart from their presuppositional properties. *Again* in (6):

(6) Mary has failed again.

does not inform us of anything apart from Mary's failing. The truth conditions of the sentence with *again* are the same as for the sentence without the particle. The existence of another occasion of failing is not asserted, but only presupposed. A third and even more puzzling characteristic is that the antecedents of some of these particles can occur in contexts that are not accessible from the position of the trigger in the sense of discourse representation theory:[4]

(7) Mary dreamt that night that she would fail the exam and indeed she did.

None of the other triggers that are central in the presupposition literature have these three properties. The only exception might be the obligatory nature of the trigger. Is the use of presupposition triggers instead of non-presupposing alternatives obligatory if the presupposition is fulfilled? I think not, but the situation is not as clear-cut as one would like. Two examples, based on the triggers *know* and *the*.

Can I say (8) when I know that *p* is the case?

(8) John believes/suspects that *p*.

Would I be pragmatically incorrect? There seems to be no problem. I merely suggest that John does not have the appropriate epistemic access to *p* to warrant the use of *know*.

If we have discussed a new girl at the office who I saw with John in town, it is again not incorrect for me to report that I saw John with a girl in town, instead of saying that I saw John with the new girl at the office: I may consider the connection irrelevant in the context. (I would suggest that they are different if the hearer would think the identity relevant.) To the extent that the standard triggers like *know* or *the* are obligatory, they are so because they are liable to mislead the hearer. Not using them can be a transgression of Grice's maxim of quantity.

The particles are different. They can only be used when the presupposition is there (since they do not accommodate) and their absence cannot really mislead the hearer if the presupposition is satisfied, since the presupposition is common knowledge already.

There are unclarities here, but it is obvious that *know* and *the* accommodate, have content and do not take inaccessible antecedents:

(9) John knew that Mary had failed.

Notice how (9) can be used to convey that Mary had failed. Knowledge is more than just belief with a presupposition and so has independent content. The truth of the presupposition is therefore not enough to make it necessary to use the word *know*.

While (10) is only acceptable under the extra assumption, that the dream is true:

(10) Mary dreamt that she would fail the exam. Bill knows that she will.

Similar examples with *the* are given in (11):

(11) a. I met the director of Peter's school.
b. Mary dreamt there was a burglar in the house. The police captured the burglar after a chase in the garden.

The first sentence can be used without Peter's school or the fact that it has a director having been mentioned before. Both facts can be unproblemtically accommodated. The second sentence in (11b) can (when it is not taken as an elaboration on the contents of Mary's dream) only be understood under the extra assumption that the dream was true.

It is clear that if we want to analyze particles as presupposition triggers, we must be able to modify our presupposition theories to make it possible for the particles to come out as a special case with special properties – that is, having no semantic content of their own, no accommodation, the possibility of inaccessible antecedents and the obligatory character of their use. I will now sketch my earlier attempt to do just that (cf. Zeevat, 2002).

2.1 Preliminaries

I am assuming a version of the presupposition theory of van der Sandt or Heim formulated in an updated semantics. In such a theory, a presupposition trigger is always added to an auxiliary information state that is introduced in order to allow the interpretation of a logical operator like negation, implication or disjunction, a modal operator or a propositional attitude. Auxiliary information states branch off from the common ground or from other auxiliary states. An auxiliary information state has access to the state from which it has branched off or to the states to which that state has access. A pronoun or a presupposition trigger in an information state *IS* can be resolved to antecedents in *IS* itself or to material in other states to which *IS* has access.

The information state that contains John's dream is not accessible from the common ground itself, as can be easily tested. But, as we saw, it is an antecedent for a particle like *indeed* (I am assuming that *indeed p* presupposes *p*). *Indeed* is very liberal in taking such antecedents:

(12) a. John dreamt that Mary would fail her exam and she failed indeed.
b. John suggested that Mary would fail her exam and she failed indeed.

But even *indeed* does not like antecedents under a logical operator, or "negative operators" like *doubt* or *deny*. This is illustrated in (13):

(13) A: If John comes, the party will be a success.
B: ??John comes indeed.
B: ??The party will indeed be a success.
A: John did not come.
B: *John came indeed/*John did indeed come.
A: John came in or Mary left.
B: ?John came indeed.
B: ?Mary left indeed.
A: Bill doubts that John will come.
B: *He will come indeed.

Possible environments are the complements of verbs like *dream, say, think, suggest* and also cases where suggestions are made indirectly, for example by saying *maybe John will come*. Iterations of these also seem to be fine:

(14) A: John said that Bill maybe has to stay home.
B: Charles also has to stay home.

Bill suggested that Mary was not pleased. She was indeed rather unhappy.

The problem is to explain why in specific cases the larger class of antecedents is not available. My explanation is that this is due to overlaps between the presupposition of the trigger and its semantic content. The presupposed complement of *know* can only be a fact, that is, information that is true in the local information state or in the information states to which it has access. So an antecedent from a dream, or from John's beliefs is not sufficient for giving the semantics of the verb *know* what it needs. The semantics of the particle *again* that presupposes an earlier occasion of the same state occurring or event happening imposes a temporal relation of precedence between the earlier state or the current one in the local information state. A state or event that is not available in the local information state cannot precede the current state or event in the information state. The cases where the weaker antecedents are possible are the ones that have little to no semantic content, that is, *too, indeed, doch/toch, wel*, and so on.

2.2 Marking principles

It is not possible to explain the obligatory occurrence of anything within the bounds of a purely interpretational theory and the presupposition theory that I have been describing is exclusively concerned with interpretation. My solution is to assume a set of marking principles in bidirectional optimality theory. Marking principles enforce the presence of certain features of the semantic input in the linguistic output. For a particle like *too* this is the presence of an item similar to the one reported in the current sentence. For accented *doch* or *toch* it is the presence of the negation of the sentence in which the particle occurs, for *indeed* the presence of the same information as reported in the current sentence. I will try to be more precise about these marking principles, when I come to speak about the context marking theory.

2.3 Non-accommodation

Blutner and Jäger (2000) reformulate presupposition theory in bidirectional optimality theory by two constraints: ***Accommodate** and **Strength**. The first constraint prefers interpretations in which accommodation does not occur, the second prefers the strongest readings. Two other constraints that we need are **Consistent** that prefers interpretations that are consistent with the context over those that are not and a constraint **Trigger** that asks that presuppositions of triggers hold in their local context. Thereby, resolutions are preferred over accommodations and if accommodations have to occur, they occur in the common ground, unless that makes the common ground inconsistent (in general, accommodating the presupposition in the common ground gives more information than adding it to a temporary information state). Within this theory one can show that adding a particle is ruled out if its presupposition leads to an accommodation. In that case there is competition with the sentence without the particle. Under the assumptions of bidirectional optimality theory the violation of the constraint ***Accommodate** is fatal for the version with the particle.

The principle is general. If a presupposition trigger has a simple non-presupposing alternative, it does not accommodate. It has been questioned whether the principle is correct for other presupposition triggers. Geurts (p.c.) has suggested that the trigger *manage* is a proper counterexample:

(15) a. John managed to open the door.
 b. John tried to open the door.
 c. John opened the door.

Here (15a) presupposes (15b), while it seems clearly in competition with (15c), and the presupposition can be accommodated.

There is however a problem with the analysis of *manage* as a presupposition trigger, presupposing that the action was tried or that it was difficult. It seems I can say (16), even if I never tried and it would not have been difficult to do so, without misleading anybody:

(16) I did not manage to phone Mary.

Manage seems to force the focus of the question on whether the action was successful or not. This in turn makes it necessary to find or construct a topic that makes it sensible to have this focus. In (16) this may be the speaker's promise to phone Mary. It is clear that if the difficulty of the action is given or an attempt to perform the action is made, that provides a suitable topic. But if the presupposition is no more than the necessity of a certain kind of topic, one should not treat *manage* as a presupposition trigger at all, but as a context marker or a speech act marker. *Manage* is an interesting case, but not a good counterexample.

2.4 Summing up

It is important to note that our alternative presupposition theory forces the replacement of **Trigger** by **Weaktrigger**, the requirement that the local context of a presupposition trigger needs to have access to the suggestion in its presupposition. But this has an unfortunate consequence. We must now also make sure that the "normal" presupposition triggers that need the full truth of their presupposition in their local context do get the normal accessible antecedents that they require. The idea that for them the presupposition itself is part of their meaning and that without the truth of their presupposition they cannot be true or false is intuitively correct, but that is not enough for a proper account. Without further constraints, we would allow updates which are only partially defined. In some of the possibilities in our information states, the presupposition is true and they can be eliminated or not depending on the truth of the semantic content. In others, however, the presupposition is not true and therefore there would be no criterion that would eliminate these possibilities or not. In sum, the update with a partially defined proposition is not properly defined. This can be remedied by a constraint **Defined**, asking us to make our interpretations an update that is fully defined with respect to the information state. But this is just a reformulation of the principle **Trigger** that we had before, with the difference that it is now limited to a subclass of the triggers, that is, those that require the local truth of their presuppositions. This shows that something has gone wrong with our attempt to understand particles as presupposition triggers. The constraint **Weaktrigger** is just a special postulate needed for particles. In fact, given that the particles do not accommodate, none of the constraints for presupposition triggers seem to play any role in understanding the particles. All the specific constraints for presupposition triggers have to do with regulating accommodation and the choice between accommodation and resolution, and, as we saw earlier on, accommodation does not play a role for particles.

There is no other possibility but to conclude that thinking of particles as presupposition triggers has no explanatory value. One can try to assimilate them to the other triggers, but it does not help in understanding particles any better. It is not inconsistent to claim that our particles are (of a kind) presupposition triggers, but the claim that this is the key to understanding particles is not tenable.

3 An alternative theory: context marking

The marking principles that we had to adopt in our analysis of the presuppositional particles were an addition to the presuppositional analysis: there is no way we can derive them from an analysis that is content with saying that they just presuppose that particular presupposition, have that particular content (if they have any).

A natural strategy towards understanding them better is therefore to turn the argument around and to investigate whether we can understand why they are like presupposition triggers if we assume that they act as markers of a relation of the content of the current sentence to the context (or to another parameter of the utterance context) which must be there because of a functional necessity (e.g., if the relation in question is unmarked, wrong interpretations result). A functional explanation is necessary for marking to be possible at all. If it were always superfluous, markers would never arise. But without a further grammaticalization process, it is not possible to understand why languages vary so much in what they have to mark. Both Dutch and English speakers can mark progressive aspect, but only English speakers have to mark it all the time. Both Russian and English can mark definitenes on NPs, but only English has to do it all the time. I will take up this issue in the conclusion.

The relations for which we have to assume marking principles are the following:

old: The content is already suggested in the common ground (*indeed, immers, doch/toch* (unaccented), *ja*).

adversativity: The content has been suggested to be false in the context (*doch/toch*, proconcessives, concessives).

correction: The content was denied in the common ground (*but, sondern*, WEL, NIET, DOCH, TOCH, DO, DIDn't).

additive: The topic has been addressed before, but the content gives an expansion of the earlier answer (*too, also, ook/auch*).

replacing additive: The topic has been addressed before, but this contribution needs to be replaced (*instead, sondern*).

contrast: The new content addresses the old topic with its polarity inverted (*but, however, maar, aber*).

Are these marking strategies universal? Empirically, this is not clear. There are many things unknown about discourse particles and they are hard to understand even in a single well-studied language. It suffices for our purposes to assume that there is a strong functional pressure to have ways of expressing these relations. That assumption is necessary, since otherwise it is not clear how we could have particles like the ones listed above or how they can appear so often. And we can indeed try to make clear what could go wrong in the interpretation process if the particles (or other forms of marking) were not there. I give my attempt to do that below.

Old marking: If an old element is not marked as old, it may be interpreted as new even if its expression is formally identical with the original introduction of the element (cf. indefinites, tense). The original element is integrated into the semantic representation by the original interpretation process, the new version will lack the connections made there.

Adversative marking: If the presence of a suggestion to the contrary is not noticed, this means that the suggestion to the contrary will be unchecked and can be the source of later errors.

Correction marking: This should lead to the retraction of the corrected element. If this does not happen, the old and wrong information may remain active. Like suggestions to the contrary, they should be marked as corrected, since otherwise they can create wrong information later on.

Additive marking: Additive marking finds an old topic and the way this was addressed before. Without the additive marking, a different topic may be assumed. Without additive marking, the two occasions of addressing the same topic remain unintegrated and can lead to wrong information due to exhaustivity effects. If one instance is noticed, it may be assumed that that is all. Or one instance may be noticed without there being a link to the other instance.

Substitution marking: Here it is essential to make sure the two ways in which the topic is addressed are kept distinct and that the two answers are not taken as a joint answer to the same topic. It is related to correction.

Contrast marking: If the polarity switch remains unmarked, it may be unnoticed. Misinterpretations can also result from interpreting the second conjunct as belonging to the topic of the first conjunct.

These motivations suggest that it is in the speaker's interest to mark these relations: without marking, she may well be misunderstood. And it is in the hearer's interest to pay attention to the marking particles since without doing that, she may well get confused.

3.1 Context marking in bidirectional optimality theory

Let us assume the convention around our particles is very simple: if the relation *R* obtains between context parameters and the current utterance, add the particle *P* to the utterance. (A more abstract version only asks for *R* to be marked somehow and so allows other marking devices apart from *P*: other particles, lexical material, constructions, intonation.) This convention (a constraint **max(R)**) overrules a constraint against special devices (an economy constraint ***Particle**). The combination of the two constraints guarantees that *P* appears if and only if *R* holds between the content and the context parameters. From the point of view of the interpreter of the utterance, an occurrence of *P* indicates that *R* holds. Since the hearer now knows the content of the utterance and already knew the context parameters, she can make sure for herself that *R* holds. This checking of *R* will force certain identifications involving the current utterance, the common ground and the topic. The check is part of the interpreter's task of reconstructing the intentions of the speaker. It is also part of the interpreter's task of integrating the new information within her overall representation of the world and doing so in an efficient way.

Can we now understand why there are similarities between presupposition triggers and a class of particles? What we have so far is a tentative explanation of two properties of our particles: the fact that they do not accommodate and the fact that their occurrence is not optional but obligatory. The other things we need to explain are the fact that they lead to a resolution process in which certain material is identified in the context and the extra embeddings under which this material may occur. The first part of this is that the relation *R* needs to be recognised as holding between the current utterance and the context parameters, and I will go through that for each *R*. The second part is just the assumption that it is the local and the not the global context (the common ground) that is relevant for finding the relations. Together, that gives the presupposition-like behaviour of the particles.

Let us go through these for each of our *R*s.

Old markers

> φ is the content of the current utterance, CG the common ground. *old*(*CG*, φ) holds iff $CG \models$ *suggested*(φ).

The relation *suggested*(φ) can be defined by a recursive definition, using a set $\{O_1, \ldots, O_n\}$ containing operators like *x dreams that, x suggests that, x believes that.*

It comes in place of our earlier inaccessible antecedents:

(17) $suggested(\varphi) \leftrightarrow \varphi \vee O_1 \text{ suggested}(\varphi) \vee \ldots \vee O_n\ suggested(\varphi)$

Each of the particles does more than just mark *R* almost by definition in this case, since repeating old information is not useful by itself. *Indeed* indicates the presence of better evidence for φ, *immers* makes φ a reason for assuming the current discourse pivot (the discourse element to which the current utterance is related by a discourse relation, normally the previous utterance), *doch*/*toch* without accent makes the old information subject of discussion again, *ja* presents it as common ground between speaker and hearer (and allows further causal or other connections based on that). This makes it hard for *immers* and *ja* to have antecedents which are merely suggested.[5]

The account of *suggested* that I give here does not take into account that suggestions are not eternal (the same holds for *normally*). If *p* has been suggested, *p* may turn out to be false after all. Or evidence may come in that makes $\neg p$ as plausible as *p*. For *normally*, it is possible to build this into the semantics, so that *normally p* can be true with respect to an information state *IS* but becomes false again on an extension of *JS* of *IS*. We can reach the same effect with *suggested p* by requiring that *may p* is a necessary condition for *suggested p*, where *may p* is the requirement on an information state that it contain possibilities in which *p* is true.[6] (18) illustrates the wrong predictions that one gets without this proviso. It should be clear that the point of adversative marking and correction is precisely to get rid of incorrect suggestions and evidence in the common ground.

(18) A: John thinks that Mary will come tonight, but he is wrong.
B: *Mary will come indeed.
B: *Susan will come too.

An amended version of *suggested* is given in (19):

(19) $suggested(\varphi) \leftrightarrow may\varphi \wedge (\varphi \vee O_1 \text{ suggested}(\varphi) \vee \ldots \vee O_n\ suggested(\varphi))$

Adversative markers

adversative(CG, φ) holds iff $CG \models$ *normally*$(\neg\varphi)$ or $CG \models$ *suggested*$(\neg\varphi)$

The semantics of *normally* is the subject of default logic and there is no standard view. I am assuming here that the truth of *normally*(p) on an information state requires that the $CG \models \psi_1, \ldots, \psi_n$ and that $\psi_1, \ldots, \psi_n$ together constitute a reason for thinking that p, while at the same time the CG must not contain a reason for thinking that $\neg p$.

The easiest case here is that of full concessives. The complement of the concessive clause gives the reason for thinking that $\neg\varphi$ and also chooses *normally* instead of *suggested*. Since the complement of the concessive connective is presupposed, it can be treated as part of the common ground. Pro-concessives (e.g., isolated *though* in English) indicate that the reason is highly activated. The other branch, based on *suggested,* is necessary. Compare (20):

(20) Mary dreamt that she failed the exam. She had passed though.

It seems impossible to construe dreams as reasons for thinking that their propositional content is true. So this is really a non-concessive adversative reading of *though*. If there is a grammaticalization path here, it goes from proper concessives to the vaguer adversative meanings.

Accented *doch*/*toch* is adversative. Partly these are pro-concessives with a normal stress (like *trotzdem, nevertheless, desondanks*), partly *doch*/*toch* has contrastive stress contrasting with an activated negative version of the current sentence. The real puzzle with *doch* and *toch* are the unaccented cases that can be proper *old*-markers without the slightest trace of adversativity. These can probably be connected to afirmation questions with a positive bias, elicited by an apparent opposite opinion of the interlocutor:

(21) A: Ich werde es ihm nächste Woche sagen.
A: I will tell him next week.
B: Dann bist du doch verreist?
B: You are away then, aren't you?

Though *doch* is here appropriate because *B* seems to imply that what *A* said is false, it also expresses that according to *B* the common ground is that *A* is abroad next week. Reanalysis as an old marker is thereby possible. Hans-Martin Gärtner (p.c.) observes that there are two intonational contours for this *doch* only one of which can be combined with the contrast marker *aber*.

An example of this use of unaccented *doch* is given in (22):

(22) Wenn er doch hier ist, kannst du ihn auch selbst fragen.
When he is here anyway, you can ask him yourself.

Corrections

correct(CG, φ) holds iff $CG \models \neg\varphi$.

The correction relation is an extreme case of adversativity: the best reason for thinking that φ is false is knowing that it is false. At the same time, unlike the weaker possibilities for adversativity, the current sentence is then not consistent with the common ground. The intended change to the common ground is a combination of retraction of (the reasons for) $\neg\varphi$ and the addition of φ as a replacement.

DOCH/*TOCH* with contrastive stress is one correction marker. Others are Dutch *WEL* and *NIET* (both with contrastive stress), English *DO* and *DOn't* (both with contrastive stress).

Additive markers

Common grounds naturally record their own history and any formal model of them must follow suit. *additive*(CG, φ) is then a combination of a complex relation to the common ground and a special intention.

The relation is between the common ground, a topic and a proposition. The topic must be such that φ addresses it. The proposition must be the strongest to hold on the common ground that addresses the topic and the common ground must "remember" that the earlier proposition addressed the topic. This calls for a special predicate:

(23) $CG \models addressed(\psi, T)$

The predicate should entail: $CG \models \psi$ and $address(\psi, T)$ and there should not be a χ such that $CG \models \chi$, $\chi \models \psi$ but $\psi \not\models \chi$ which also addresses T.

On a proper model of topic, addressing should be a formal relation between the formal topic and the sentence. For example, on a model of topics where they are equated with Hamblin-style questions (Hamblin, 1973), a proposition addresses a topic iff it is a member of the topic.

The intention of the speaker is that now the conjunction of ψ and φ becomes the information that the common ground has about the topic. That is, *addressed*(ψ, T) will be false on the new common ground and *addressed*($\varphi \wedge \psi$, T) will be true. Close in functionality to *additive* markers are "*other*- markers", like *another* in *Another girl walked in*. If we think of the noun *girl* as a topic that is addressed by the indefinite, their treatment is formally the same. But I am not sure it makes sense to think of the noun semantics as an additional topic *which girl*?.

Replacing additive markers

Replacing additive markers like *instead* are only different in the intention with which they are used and place the same condition on the context. We want to ensure that the proposition that addressed our topic before is replaced by the current proposition φ so that afterwards, the common ground has it that *addressed*(φ, T) is true and *addressed*(ψ, T) is false.

The choice between additive and replacing additive markers explains the relative uncomfortability of antecedents that are only suggested for these markers. Example (24a) suggests that Sue is in Spain next to John, the second suggests that the dream is false. Leaving out the particle completely is not an improvement. We now no longer mark that the topic has been addressed before:

(24) a. Mary dreamt that John is in Spain. (?) Sue is also in Spain.
b. Mary dreamt that John is in Spain. (?) Sue is in Spain instead.
c. Mary dreamt that John is in Spain. (?) Sue is in Spain.

(25) illustrates how subtle this is. The situation (A and B are children in a secret phone call) makes it clear that B's parents do not know about the other child. And many people find the example mildly anomalous:

(25) A: My parents think that I am in bed.
B: My parents think that I am also in bed.

One way of explaining the anomaly is therefore the assumption that *too* and *instead* are not pure context markers, but also speech act markers for the specialized speech act of adding to/substituting information in an old topic. In our last two examples this function is not applicable.

Contrast markers

The most complicated relation I consider here is that of contrast and one might well wonder whether it belongs in this sequence. I think it does and that it is a mere coincidence that contrastive markers often appear as coordinating sentence connectives. In German, *aber* (*but*) also appears in later positions in the sentence and an extensive corpus study by Schösler (2002) reveals that there is no essential difference in these uses, which are translatable by *echter* in Dutch or by *however* in English. My provisional analysis, derived from Umbach (2001), goes as follows, using the machinery I introduced above.

Let ψ be the discourse pivot (the predecessor of the current utterance) and let $CG \models addressed(\psi, T)$. φ is contrastive iff it directly or indirectly addresses *negate*(T). Here, *negate*(T) is the topic that is addressed by the negation of any formula that addresses T. Using the conception of topic derived from Hamblin (1973), we can obtain *negate*(T) from T by replacing all T's elements by their negations.

> The sentence S indirectly addresses a topic T iff the common ground updated with the information that S answers its own topic Q entails an element of the topic T.

I illustrate the analysis by (26). In (26a) the second conjunct directly addresses the topic of the first sentence: *Who was ill?*. I assume that in (26b) and (26c) this is also the topic of the first conjunct. In (26b) we can construct the topic of the second sentence as, for example, *Who was as fit as a fiddle?* or *Was John as fit as a fiddle?* In both cases the answer entails that John was not ill, which directly addresses the negation of the topic of the first clause. In (26c) the topic of the second conjunct is something like: *What about John?* The fact that the answer does not include that he was ill, together with the fact that the negation of the topic of the first conjunct must be addressed, implies that John was not ill:

(26) a. Mary was ill, but John was not.
b. Mary was ill, but John was as fit as a fiddle.
c. Mary was ill, but John came to the party.

This last type of inference is typical for contrast. In (27), we can infer that Mary did not attend the party, even though world knowledge tells us that many people wash their hair before going to a party:

(27) John went to the party, but Mary washed her hair.

With Umbach, I hold that the concessive uses of contrastive markers are derived uses.[7] (28a) can be paraphrased as (28b):

(28) a. Although Mary was ill, John went to the party.
b. Mary was ill, but John went to the party.

Here *but* functions as a pro-concessive, taking its antecedent from the first conjunct. This requires that the common ground makes Mary's illness a reason for thinking that John would not go to the party (it may be known that in such cases he feels his duty is at home). Perhaps the reanalysis is based on the fact that often one positive answer to a topic makes further positive answers more plausible. If you know Mary and John, the fact that Mary goes to the party can make it much more plausible that John will go there as well. Where this is so, contrastive *but* in (29) also marks adversativity. Notice that extra adversative markers seem out of place:

(29) Mary is going to the party, but John is not.

A simple treatment of *and* along the same lines (as a *topic maintenance marker*) is to say that *and* forces the second conjunct to at least indirectly address the same topic. This is the essence of the analysis given by Gomez-Txurruka (to appear).

4 Conclusion

I have discussed so far what context marking is if we assume that syntax tells us to mark certain relations of the current utterance to context parameters like topic and common ground and if the interpreter's task is just to reconstruct the speaker intention. We have assumed that the presence of context markers is largely explainable by the difficulties facing the hearer in properly integrating the current utterance with the information that she has already got. Particles in this view are the signals from one copy of the human conversational faculty to another. They may not make much sense to us as rational agents, but they do a lot for the proper storing and connecting of the bits and pieces that come in.

The only assumption that we need to make for obtaining the presuppositional behaviour of some of the particles that I discussed is the assumption that for embedded occurrences of triggers, the local context is the one with respect to which marking needs to take place. This will explain those cases in which the common ground does not itself have the required relation to the content of the sentence, as in (30):

(30) Falls du nach Berlin kommst, triffst du ihn ja.
In case you come to Berlin, you will meet him *ja*.

The presuppositional character of some of the particles is basically the reconstruction by the hearer of the relation marked by the particle under which the utterance is made. This forces the identification of a topic or a proposition in the common ground. There is no accommodation because the relations are overt. It makes no sense to warn the hearer about a relation that does not obtain. Suggestions can open topics and address them positively and negatively. That is enough to understand why old, adversative and additive markers can take indirect antecedents.

It is therefore unnecessary to invoke "presupposition theory" for the analysis of discourse particles. In fact, one may wonder whether presuppositions – or presupposition triggers – must be considered a natural class in linguistics, a category that explanations can be based on. After all, the triggers normally considered in the presupposition literature fall into at least three classes: the ones considered here, the referential devices like names, definite descriptions, clefts, and so on, and the lexical presupposition triggers (the largest and least studied group, including next to *bachelor* and factives, most adjectives, nouns and verbs.). This chapter should have convinced you that there are serious differences between the particles and the other triggers. Zeevat (1992) discusses differences in their projecting behaviour between the lexical and referential triggers.

An attempt to understand particles as presupposition triggers also runs into the problem that many are not. It is clearly the case that more particles can be analyzed as context markers, but this should not fool us into thinking that context marking is all there is to particles. Very obviously, a great many discourse particles mark speech acts. The clearest case are markers like Chinese *ma* that makes yes-no questions out of assertions as in (31):

(31) Ni hao ma?
You good QUESTION-PARTICLE.
Are you OK?

Or take the unaccented *wel* in Dutch as in (32):

(32) Het komt wel goed.
Don't worry.

The particle tones down the preconditions of normal assertion (the speaker has to believe she knows what she is telling the hearer) to mere "trust me" belief. This – like a repetition or a correction – is a specialization of the speech act of assertion. The context markers we considered before also have aspects that relate them to the evidentiality dimension of speech acts: *indeed* also indicates an increase in evidentiality with respect to the antecedent, accented *doch* can have a similar function, indicating that there is now evidence that what we thought before to be false has now turned out to be true.

I will not attempt a formal theory of speech acts in this chapter, but just give an outline. We assume that there are at least three dimensions. The first dimension is the set of preconditions for the speech act: what must be the case with the context of the utterance that makes it possible to carry out the speech act. The second dimension is the aim that the speaker wants to achieve with her speech act. The third is the effects that the speaker achieves with the speech act independently of whether she properly reaches her aim. (Her speech act must still be recognised as such by the interlocutor, but the interlocutor does not give the intended response.)

Context marking is in the first dimension: it preconditions the speech act and changes the defaults assumed there. We have seen that the two varieties of additive marking (*too* versus *instead*) also affect the second and third dimension. With *too*, we intend to bind an old topic question to a new value that is obtained by adding the value specified in the sentence to the old value. With *instead*, we intend to replace the old value by the value specified in the sentence. This also affects the third dimension: in the case of *too*, the speaker endorses the old value of the topic in addition to the value specified in the sentence, whereas, in the case of *instead*, she disagrees with the old value and only expresses her belief that she knows that the value is as expressed in the sentence.

Assertions have the following preconditions, intentions and minimal effects. It is tempting to think that all other speech acts derive from this notion of standard assertion by overriding some of the default settings, by using marked sentence forms, intonation, particles, and so on:

> **Assertion**: *p*
>
> **Preconditions**: the common ground contains no reason for thinking that *p* is true or false, the hearer wants to know the answer to a new topic *Q*
>
> *p* settles *Q*
>
> Intention: that it become common ground that *p*
>
> **Minimal effect**: to make it common ground that the speaker believes she knows that *p*

Adversative and old-marking changes the preconditions. We obtain corrections and reconfirmations when the adverse or old information is in the common ground itself. Additive markers make the topic question old. Other markers (*wel, maybe, schon*) change the operator under which the new information enters the common ground from *the speaker believes she knows* to weaker ones: *the speaker thinks it is probable that, the speaker thinks there is a chance that* or *the speaker thinks that*. The effect of an accepted weakened assertion of this kind is also changed: *it is probable that, there is a chance that*, speaker and hearer *think that*. Markers are possible that change the intention and minimal effect entirely, like the tag *isn't it* (the speaker believes she knows that *p*, but wants to know if the hearer agrees instead of proposing it as common ground directly), the Chinese particle *ma* (the speaker wants the hearer to decide between *p* or ¬*p*), the intention is that *p* or ¬*p* are added to the common ground. Yet other markers (performative verbs, *please*) turn the assertion into promises or requests.

In promises and requests, the intention and the preconditions are the same (or can be thought of as the same since *p* becomes a part of the common ground after the promise or request is accepted, not as what the interlocutors believe they know, but as something they agree will be brought about by them). But the minimal effect is different: the speaker wants *p* to be true.

The speech act that is most different is the *Wh*-question. Here the precondition is that the common ground does not settle the question yet, the intended effect is that the common ground settle the question, and the minimal effect is that the speaker wants the common ground to settle the question. A proper treatment of this requires delving into the semantics and pragmatics of *Wh*-questions, something that I want to defer to another occasion.

It is clear that default settings for speech acts can be found by considering what goes on with the most unmarked surface forms. One can study the default assertion, the default question, the default request, the default acknowledgment, and so on. One can then add more and more marking and obtain an ever-increasing range of speech acts. There are two interesting questions here. One is what the semantic type of a particle (intonation pattern, syntactic inversion, etc.) is. If I am right, they map speech acts into speech acts and we need the type of a speech act for a proper mathematicized account. In the more pedestrian environment of unification-based semantics, we have default unification as a standard tool, however: the particle is a slot filler that can override default settings. The second question is whether we are dealing here with universal defaults and universal marking patterns. Empirically, this is a hard question. As a Dutch speaker who occasionally attempts to speak English, German and Italian, I can only say that sometimes one needs to considerably change the speech act and its propositional content in order to achieve roughly the same effect.

There is a principled question as well here. If we assume with the founders of optimality theory that constraints are universal, then our marking constraints must be universal as well. They can of course be unsatisfiable if the language lacks appropriate marking devices or they can be outranked by an economy constraint that prohibits marking.

To take up our earlier example, Dutch[8] can mark progressive aspect but does it optionally, whereas English does it obligatorily; Russian can mark definiteness but does it optionally, whereas in West European languages it is obligatory. Within OT, we must say that **max(progressive)** or **max(definite)** is outranked by ***Structure**, in Dutch and Russian respectively. This is not the whole story, since we must allow optional marking. I assume that bidirectionality is responsible for this: if the speaker can see that the unmarked version leads to the wrong interpretation in the interlocutor, marking is necessary, even if it transgresses ***Structure**.

In a theory of speech acts in which there are default settings for various parameters, it is possible to understand better why the use of particles is governed by marking principles. If the particle is not there to override the default setting, the default is assumed. This means that the speaker will be misunderstood if she omits the particle when she intends a speech act with a property expressed by the particle. Obligatory marking can therefore arise from the zero situation with an ambiguous speech act form. If marking has become possible, if there is a statistically based preference for one of the readings (possibly resulting from the optional marking) and if there are sufficiently many misunderstandings, Jäger's Bidirectional Learning Algorithm (Chapter 11) lets the optionality of the marking decrease and the bias for misunderstanding the unmarked form as standard increase. This promotes further marking and stronger bias towards misinterpretation of the standard form. Under the appropriate statistical conditions this leads to obligatory marking.

It is not necessary to assume universal marking principles or universal default settings. The possibility of marking may create a bias towards the opposite interpretation which in turn may create the default setting and the obligation to mark.

It should be clear that a proper account of speech act marking needs a lot of further elaboration. But the concept of a speech act semantics as a successor to dynamic semantics seems the most promising direction in which a fully formal approach to the semantics and pragmatics of particles could be achieved.

Acknowledgments. Thanks go to Nicholas Asher, Kerstin Fischer, Reinhard Blutner, Janneke Huitink and two anonymous reviewers for helpful suggestions and corrections. All errors are my own.

Notes

1. I thank Nick Asher (p.c.) for this argument.
2. This claim needs a number of provisos. First of all, partial accommodation does not seem a problem, as is made quite clear in Kamp and Rossdeutscher's (1994) treatment of *wieder*. Also, the non-linguistic context can provide salient antecedents (*I watched too*, the morning after the world cup final). Finally, there are counterexamples which involve the speaker or the hearer as in: *Do you want a beer too?*
3. Corpus studies by Tim Kliphuis and myself suggest that omitting them nearly always leads to awkwardness, or to differences in the implicatures.
4. A subordinate context is inaccessible at a position *x* iff the information that it contains is not entailed at *x*. A subordinate antecedent for a pronoun occurring at *x* is inaccessible if the existence of the antecedent is not entailed at *x*.
5. This makes a proper account of them dependent on the constraint **Defined** that I discussed earlier on.
6. Thanks to Marie Nilsenova and Robert van Rooy for pointing out this complication and to Manfred Krifka for the example.
7. This can be doubted. Prof. Asiatini of the Tblisi State University noticed (p.c.) that in Georgian the concessive and contrastive uses of *but* are lexicalized in a different way. This shows at least that normal language users do not conflate the two uses and that contrastive markers do not always allow concessive interpretations.
8. Certain Flemish dialects, including the Antwerp dialect, are an exception and pattern like English.

6 Input–Output Mismatches in Optimality Theory

David Beaver and Hanjung Lee

1 Introduction

Bidirectional Optimality Theory allows us to see a wide range of problems which would previously have been considered unrelated from a new perspective, the perspective of asymmetric relationships between input and output. For interpretation, the input is a form and the output a meaning, and for production the input is a meaning and the output is a form. A mismatch is any case where there is no isomorphism between the space of meanings and the space of forms, say because one form has no meaning, or multiple meanings, or because a meaning is inexpressible, or may be expressed in multiple ways.

Is there such a thing as a perfect language, one that would lack any mismatch of this sort? Certainly, there are subsystems of natural and formal languages that, if taken in isolation, would be perfectly symmetric. For example, the Arabic notation for integers (assuming that initial zeroes are ill-formed) stands in a one to one relationship with the abstract semantic space of integers. But even formal languages are commonly not perfect in this very strong sense. For example, in first order logic there may be multiple constants referring to the same individual, and more generally there are an infinite number of ways of expressing any proposition that can be expressed at all. There may also be objects in the model for which there is no corresponding constant, or facts that are true in a given model or frame and yet inexpressible in first order logic. As far as form–meaning symmetry goes, the only way that first order logic scores qualitatively over natural language is that the former is (when properly notated, and interpreted with respect to a specific model) unambiguous: for any form there is exactly one meaning.

Along with ambiguity, we will be considering optionality, ineffability, uninterpretability, blocking and freezing. All of these involve a mismatch between form and meaning, and we will study how various versions of OT handle these mismatches.

Initially, we will be considering simpler, relatively standard OT architectures. The first two of these are unidirectional. What we will term *naive* OT *production* is the approach seen in most OT syntax papers, and is close to the model that is used in OT phonology. To recap what we assume is already familiar to most readers of this article, naive OT production starts with some representation of meaning as input, and a set of candidate outputs provided by a function referred to as GEN. A set of linearly ranked constraints is then used to select between candidate surface forms. The second unidirectional approach, not surprisingly, works the other way: we will term it *naive* OT *comprehension*, although Hendriks and de Hoop (2001) term it OT *semantics*. The input is a surface form, GEN offers a set of candidate meanings, and the linearly ranked constraint set is used to find the best meaning for the given form.

In this chapter we are not concerned with processing issues, computational complexity or the psychological plausibility of the OT tableau method. Rather, we take an abstract view of the languages that various OT models generate. As a result, and despite the danger of terminological confusion, naive OT production can be considered a theory of both comprehension and production. The same goes for naive OT comprehension. The reason is that both unidirectional accounts ultimately capture a relation between meaning and form, or, equivalently, a set of meaning–form pairs. Thus, naive OT production characterizes a language as the set of pairs of meanings and forms such that for the given meaning, the form is optimal. Likewise, naive OT comprehension characterizes a language as the set of pairs of meanings and forms such that for the given form, the meaning is optimal.

Some OT architectures provide grammars that cannot be reduced to a set of meaning–form pairs. One of these, which we will term *naive back-and-forth* OT, consists of an obvious combination of naive OT production and comprehension: the first is used for production only, and the second for comprehension only, an architecture discussed by Hendriks and de Hoop (2001). Note that even if the constraints used in each direction are the same, this model may not assign a consistent relation between meanings and forms. In particular for some choices of constraints, if you take a meaning, apply naive OT production to get a form, and then apply naive OT comprehension, you may not get back to the original meaning.

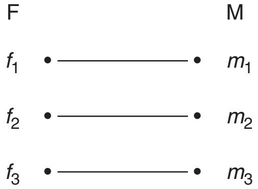
In addition to these three *naive* models, we will also consider four more sophisticated variants, sophisticated in the sense that they have been specifically designed to target some of the mismatch phenomena we will be discussing. The four other models to be studied are the *strong bidirectional* OT and *weak bidirectional* OT of Blutner (2000), and the *asymmetric* OT models of Wilson (2001) and Zeevat (2000). We will introduce these models individually later in the chapter.

2 Patterns of mismatch

In this section we will consider various phenomena involving mismatches between form and meaning, and discuss the significance of these phenomena for naive OT architectures.

Perfect language

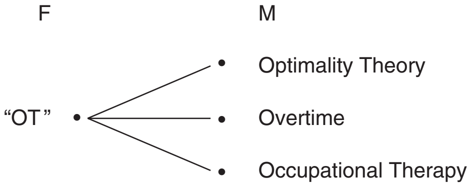
Before considering the 'imperfections' of natural languages, let us briefly gaze upon perfection. A perfect language would be one in which there was a one-to-one correspondence between forms and meanings:



As noted, even formal languages usually fail to achieve this level of perfection.

2.1 Ambiguity

This is the case of multiple meanings corresponding to a single form.[1] An example is the multiple interpretations of the abbreviated form "OT":[2]



As regards unidirectional OT models, ambiguity constitutes a *prima facie* problem for naive OT comprehension, but not for naive OT production.

In principle, a given constraint set may produce multiple outputs for a given input. Thus, there is potential for modeling ambiguity in OT comprehension. However, in practice the multiple outputs of a linearly ranked constraint set do not provide a good tool for modeling natural language ambiguity. The problem can be seen as follows: although the constraints are merely preferences, there is no way to distinguish in the output set between winners that result from strong preferences (i.e., highly ranked constraints) and winners that result from weak preferences (low ranked constraints). As a result, interpretations which one might expect to be available, if mildly dispreferred, end up being ruled out altogether.

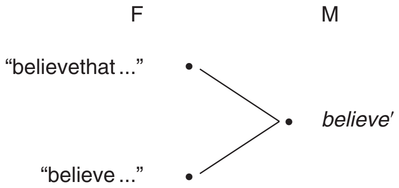
Standard examples are found in phonology. For instance, consider the neutralization between “d” and “t” in standard Dutch and English. In Dutch, “rat” (“*rat*”) and “rad” (“*wheel*”) may be pronounced identically, as discussed by Boersma (1998) and Hale and Reiss (1998) and also by Zeevat (2000), and the same goes for “wader” and “waiter” in many US varieties of English. Suppose we have the spoken Dutch input /rat/. By assumption, there is a faithfulness constraint preferring interpretation via the underlying phonological form [rat] to interpretation via underlying [rad]. If we assume linear ranking of constraints, then this faithfulness constraint is either dominated by a constraint preferring the reverse interpretation, or it is not dominated by such a constraint. Either way, /rat/ comes out unambiguous. Similarly, for US English phonetic-phonological faithfulness would lead us to expect unambiguous interpretation of /wɵɪɾə/ as something which wades. But in fact both this and the alternative interpretation, as someone who waits on tables, are available. For other examples of why ambiguity is problematic for unidirectional OT, the reader is referred to Anttila and Fong (2000) and Asudeh (2001).

For naive production, ambiguity presents no obvious problem. While unidirectional OT tends to mitigate against multiple outputs for a given input, it actually favors multiple inputs producing the same output. The /rat/ example could be derived if some constraint favoring devoicing in the given phonological environment outranked the constraint enforcing voicing faithfulness. In that case, both /rat/ and /rad/ would be realized as [rat].

What does naive back-and-forth OT predict? As regards production, the ambiguity is correctly predicted, but comprehension examples like those above are problematic: no ambiguity is predicted.

2.2 Optionality

Here we have multiple forms corresponding to a single meaning. Note that some use *optionality*[3] to describe cases where a word or expression may be added to a given form without apparent meaning change, as for example in the often claimed optionality of the complementizer “that” in English propositional complements.[4]



Synonymy, as opposed to *optionality*, is often used to describe semantic identity of two otherwise unrelated expressions, as in a case of lexical synonymy. For example it might be claimed that "creek" and "brook" are synonyms. For our purposes *optionality* and *synonymy* are not differentiated.

There is a further issue of whether true synonymy or optionality ever occurs in natural language: Bolinger and others have argued that any difference in form must correspond to a difference in meaning, where *meaning* is understood broadly to include register effects, subtle sociological connotations or other pragmatic significance.

A classic case of optionality is that of so-called *free* word order languages, even though variation of word order typically has information structural significance. Consider subject–object NP ordering for Korean transitives. For canonical Korean transitives, case marking distinguishes the subject from the object: both OSV and SOV orders are possible, but word order does not determine argument role. Here we may say there is optionality in word order, but it must be borne in mind that in Korean the choice between OSV and SOV is related to the relative information status of the subject and object, so we can talk of optionality relative only to a concept of meaning that excludes information status.

Optionality being, from our abstract perspective, the reverse of ambiguity, it is easy to see how the naive OT models fare. Optionality is unproblematic for naive comprehension OT, but is problematic for naive production and naive back-and-forth OT.

2.3 Ineffability

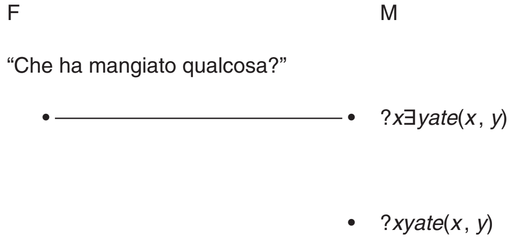
In standard OT there is always at least one winner. So whatever meaning is used as the input, standard OT grammars predict an output. By far the majority of OT grammars only describe single clauses, or relatively simple clause combinations. Thus for any meaning given as input, a relatively simple sentence is produced as the output form. In many cases this has proven problematic.

Consider the case of Italian *Wh*-questions. In Italian, multiple *Wh*-questions are infelicitous for most speakers, yet an OT grammar of Italian would presumably produce an output when given an input corresponding to the meaning of an English multiple *Wh*-question. So while in English the input meaning that we gloss as in (1d) might be realized as in (1a), in Italian the analogous form (1b) is infelicitous. An OT grammar of Italian may then, as Zeevat (2000) speculates, produce a form like that in (1c) for this input. This is a felicitous sentence, but not appropriate for the given input, since it would be interpreted as in (1e):

(1) a. Who ate what?
b. *Che ha mangiato che cosa?
Who has eaten which thing
'Who ate what?'
c. Che ha mangiato qualcosa?
Who has eaten something
'Who ate something?'

'*Who ate what?'
d. $?xy\mathit{ate}(x, y)$
e. $?x\exists y\mathit{ate}(x, y)$

In diagrammatic form, the mismatch appears as an unconnected node in the space of meanings:



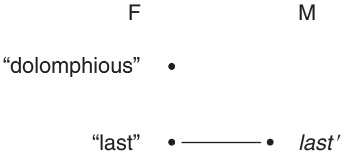
Ineffability presents a problem for naive production OT and naive back-and-forth OT. By assumption, the nature of the input (meaning) should not vary cross-linguistically, so the range of licit inputs is the same for English as for Italian. And in unidirectional OT any input produces some output, so there should be no such thing as ineffability. This is not a problem that could be wriggled out of using clever choices of constraints or a special approach to ranking. No, if naive production OT is to be taken seriously as a model, then the very existence of ineffability would have to be denied. We would have to claim that every input has an output, and perhaps broaden GEN to include multiple sentence outputs combined with appropriate gestures amongst the candidates. This would model an Italian expressing the meaning of a multiple *Wh*-question via a complex discourse and, to use a common stereotype, plenty of hand-waving. We will not pursue this line of thought further here, but assume, in agreement with, for example, Fanslow and Féry (to appear) and Zeevat (2000), that ineffability does occur, and that our model of grammar must account for it.

For naive comprehension OT, ineffability is no problem at all. While every form corresponds to some meaning in this model, there is no reason at all why all meanings should correspond to some form.

2.4 Uninterpretability

The inverse of ineffability is uninterpretability, a form with no corresponding meaning.[5] Thus Chomsky maintains that "colorless green ideas sleep furiously" is grammatically well formed, but lacks any semantic interpretation. In Edward Lear's nonsense poem *The Owl and the Pussycat,* "runcible" lacked conventional meaning when he applied it to "spoon", and still lacks conventional interpretation in its application to "cat", unless it is a cat that is curved like a spoon and has three prongs, one with a sharp edge. Lear's "dolomphious", an adjective of ducks, still lacks conventionalized meaning.

We have the following type of picture:[6]

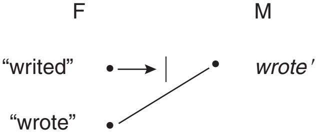


By obvious analogy with the case of ineffability, the existence of uninterpretable strings is problematic for naive comprehension OT and for naive back-and-forth OT, since they will provide an interpretation for any string given as an input. Uninterpretability is unproblematic for naive production OT.

2.6 Blocking

Blocking is a process which prevents or removes asymmetries. The most common example cited is that where a given meaning could potentially be realized either by an idiosyncratic irregular form, or by a regular productive morphological process applied to a root. The existence of an irregular form may then be said to block the regular form:

(2) a. wrote, *writed
    b. sheep [**+pl**], *sheeps



The existence of a lexical form produced by semi-productive morphology may also block a phrasal form. Poser (1992) and Bresnan (2001a) consider English comparative and superlative adjectival inflections: the existence of "cheaper" can be said to block "more cheap" in (3), whereas the absence of "expensiver" means that "more expensive" is available. Note that from a purely logical point of view, we could analyze "more expensive" as blocking "expensiver", but it is standard to analyze simpler forms (e.g., a single lexeme) as blocking more complex ones rather than the other way around:

(3) a. cheaper/cheapest, ?more/?most cheap
    b. *expensiver/*expensivest, more/most expensive

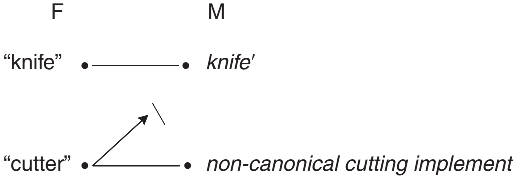
From our birds-eye perspective, we would equally term as *blocking* a case where the existence of a special meaning prevents an otherwise logically possible interpretation. Idiomatic meaning may be of this sort: "Mary kicked the bucket" could mean just that, but is invariably interpreted less fortunately. We can also understand cases involving alternative binding possibilities for pro-forms in terms of blocking of meaning (Levinson, 2000; cf. Huang, 2000). For example, in the Marathi case in (4) a preference for more local anaphora resolution prevents resolution outside of the clause:

(4) Tom$_i$ mhanat hota [ki Sue$_j$ ni *swataahlaa*$_{*i/j}$ maarle]. [Marathi]
Tom said that Sue ERG ANAPHOR-ACC hit
'Tom said that Sue hit herself/*him.' (Dalrymple, 1993, pp. 19–20)

Note that none of the naive OT models provide any account of blocking, or of the variant *partial blocking* to which we now turn.

2.6 Partial blocking

Blocking can leave a form unemployed, but the unemployed form may soon find a new job, generally expressing something closely related to but subtly different from the canonical interpretation that one might have expected. This is partial blocking: an asymmetry is eliminated, but removal of a link creates a new form–meaning pair. An example from Kiparsky (1983) is the interpretation of "cutter", a nominalization involving application of a regular and productive rule ("-er" addition). The observation is that when someone refers to "a cutter" they could not ordinarily be referring to an object for which a standard idiosyncratic expression exists, like "scissors" or "a bread knife". So "a cutter" is interpreted as a non-canonical instrument used for cutting:

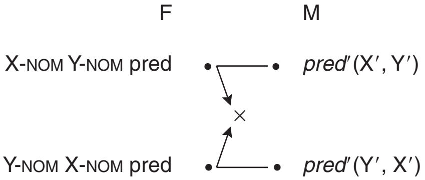


Similarly, it has often been argued that the existence of a lexical item "kill" blocks "cause to die" from having its canonical meaning, that is, the meaning that would be derived compositionally. "Cause to die" comes to denote a non-canonical killing, for instance one where the chain of causation is unusually long or unforseeable (cf. McCawley, 1978).

There are also cases where a form–meaning pair is blocked because the form has a different interpretation, and so the meaning comes to be expressed in another way. For example, "computer", "calculator" and "reckoner" are all understood to refer to non-humans, but originally referred to humans who computed, calculated or reckoned. When we wish to refer to a human who performs these tasks, or one who performs them particularly well, we now use terms like "human calculator", which once would have been tautological.

2.7 Freezing

Freezing is a phenomenon which can be seen in terms of a combination of ambiguity and optionality: it may constrain optionality to prevent ambiguity. Above, we mentioned word order freedom for the arguments of canonical Korean transitives. The caveat *canonical* is crucial, since the optionality vanishes for certain classes of verbs, notably a group of psychological predicates. For these predicates the subject and the object have identical case marking, in fact nominative case. This identity of case marking has the potential to create ambiguity, since one cannot tell from the morphological form alone which is the subject and which is the object. For verbs in this class, but for no others, word order is the primary means used to represent argument structure, with SOV order fixed in most contexts. In this case, if we may speak teleologically, it appears that word order has been frozen in order to prevent ambiguity of argument structure. Graphically, we may represent the situation, in which multiple input–output mismatches are simultaneously blocked, as follows:



As was the case for blocking, freezing phenomena are not modeled by any of the naive OT strategies. However, we will now turn to a more detailed consideration of a class of bidirectional OT models which were originally introduced precisely because they suggested a line of attack for such phenomena.

3 Strong bidirectional optimization

Besides the phenomena of form–meaning mismatches we discuss here, arguments for bidirectional optimization have come from various sources. These include the production/comprehension asymmetry in child grammar (Smolensky, 1996), decidability in computational processing (Kuhn, 2001a) and learning algorithms (Jäger, this volume). Given that production-based and interpretation-based optimization are both well motivated, a question immediately arises as to how the two directions of optimization can be combined into a coherent theory of language structure and interpretation.[7] One option is to combine them conjunctively, producing a model which Blutner (2000) calls the *strong bidirectional* OT model (this will be compared with a *weak* version in Section 4). The idea is that in order to be grammatical, a form–meaning pair $\langle f, m\rangle$ has to be optimal in both directions of optimization. That is, a form–meaning pair is strong OT optimal iff the form produces the meaning in Interpretation OT and the meaning produces the form in Production OT. So we arrive at the following definition of bidirectional optimality (The connective "$>$" is read as "more harmonic than" or "more economical than"):

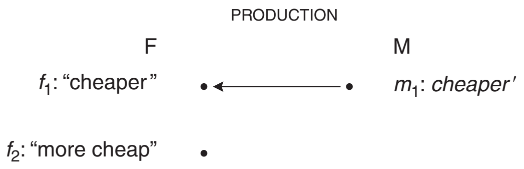
(5) $\langle f, m\rangle$ is strong OT optimal iff:

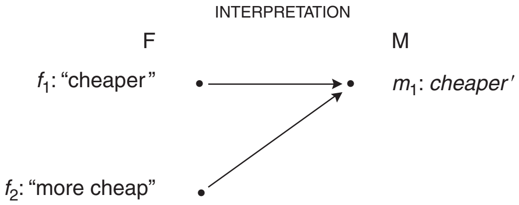
a. $\langle f, m\rangle \in$ GEN,
b. there is no $\langle f', m\rangle \in$ GEN such that $\langle f', m\rangle > \langle f, m\rangle$, and
c. there is no $\langle f, m'\rangle \in$ GEN such that $\langle f, m'\rangle > \langle f, m\rangle$.

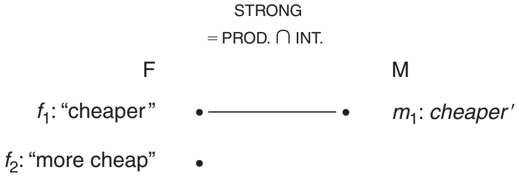
For a more detailed discussion of the formal properties of this notion of optimality, the reader is referred to Blutner (2000) and Jäger (2002).

Strong OT removes form–meaning pairs that are only optimal under one direction. In this way, it produces strictly fewer form–meaning pairs than either naive production or interpretation OT would with the same constraint ranking, and consequently it can model both ineffability and uninterpretability. Ineffability results if the optimal realization for m is the surface string f, but in comprehension-based optimization for f we get a different meaning m' ($m \neq m'$). So, m' blocks m, making m ineffable. Uninterpretability occurs when the interpretation-based winner m for the form f has a different form f' in production-based optimization. See Section 4 for a more detailed discussion and illustration.

Strong OT offers a treatment of synonymy blocking, a phenomenon which remains unaccounted for in (unidirectional) interpretation OT. Suppose that we are analyzing two forms f_1 and f_2 which are semantically equivalent and that we have some meaning m_1 that is optimal for both forms. In Interpretation OT the two forms would not belong to the same candidate set and thus would both be grammatical. In the Strong OT model, f_2, even if optimal in the interpretation-based optimization, may be blocked by the more economical alternative form f_1. Hence, the form–meaning pair $\langle f_2, m_1\rangle$ is removed from the set of the language generated by the Strong OT system. We can illustrate this with the following picture:



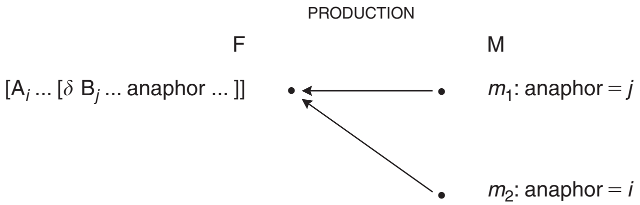


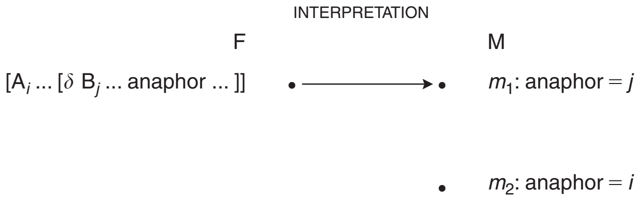


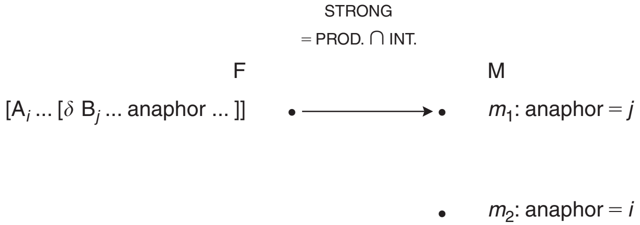
Strong OT also opens up a simple way of modeling the blocking of meaning, a phenomenon which is unaccounted for under unidirectional production OT. Consider the Marathi example from Section 2 repeated in (6) below:

(6) Tom$_i$ mhanat hota [ki Sue$_j$ ni *swataahlaa*$_{*i/j}$ maarle]. [Marathi]
Tom said that Sue ERG ANAPHOR-ACC hit
'Tom said that Sue hit herself/*him.'

Example (6) has the form [A$_i$ … [δ B$_j$ … anaphor …]], in which A and B are potential antecedents for the anaphor and δ is the domain in which the anaphor must have an antecedent (the minimal finite clause that contains the anaphor). Parsing this sentence will result in two classes of analyses: one in which the binding relation is local (i.e., anaphor = *j*) and one in which the binding relation is non-local (i.e., anaphor = *i*). In production-based optimization, the two interpretations do not compete with each other and thus the sentence is grammatical for both interpretations. In interpretation-based optimization, the former interpretation is preferred to the latter interpretation by a locality constraint on binding. As a result, anaphora resolution outside the clause is blocked by local anaphora resolution and hence removed from the set of interpretations generated by the Strong OT system. Taking together the two directions of optimization, we correctly predict not only that (6) is interpreted as *say(Tom,hit(Sue,Sue))*, but that it is the preferred way of expressing this meaning:







Strong OT also provides a solution to the problem of freezing: Lee (2001a) presents an OT treatment of word order freezing based on such a bidirectional optimization.[8] As discussed in Section 2, Korean (non-agentive) psychological verbs take two arguments bearing nominative case. For these verbs, object-subject order is not possible (without very strong contextual licensing):

(7) Mary-ka tokile kyosa-ka philyoha-ta. [Korean]
Mary-NOM German teacher-NOM need-DECL

'Mary needs a German teacher.'
*'The/a German teacher needs Mary.'

If the order of the two nominative arguments in (7) is switched as in (8), the interpretation is switched too:

(8) Tokile kyosa-ka Mary-ka philyoha-ta. [Korean]
German teacher-NOM Mary-NOM need-DECL

'The/a German teacher needs Mary.'
*'Mary needs a German teacher.'

In contrast, the argument NPs of canonical transitive verbs can appear in either order preceding the verb, and change in their order does not change the basic meaning of the sentence:

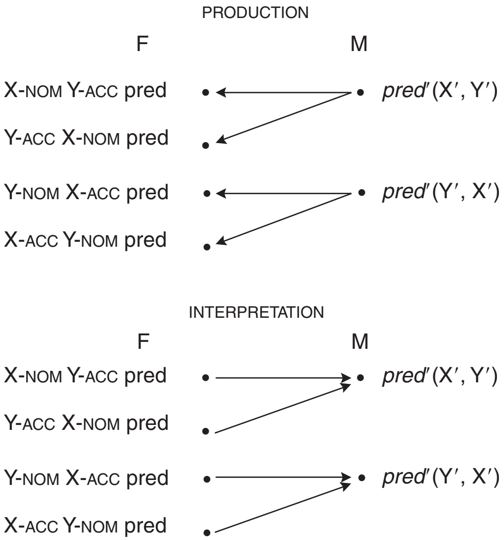
(9) a. Mary-ka nonmwun-ul sse-ss-ta. [Korean]
Mary-NOM paper-ACC wrote-PST-DECL
'Mary wrote a paper.'
b. nonmwun-ul Mary-ka sse-ss-ta.

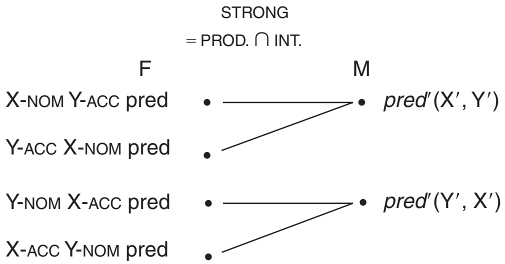
Lee (2001a) assumes two conflicting constraints on word order first proposed by Choi (1999); a canonical word order constraint (10a) and a discourse-based word order constraint(10b):

(10) a. SO: Subject precedes object.
b. TOPIC: Topic precedes non-topic.

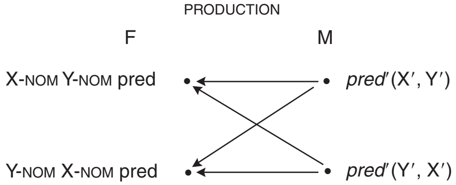
The ranking TOPIC ≫ SO ensures that object-subject order is optimal, if the object is marked [+TOPIC] in the input. When it is not marked [+TOPIC], however, the TOPIC constraint is vacuously satisfied and the lower-ranked SO constraint becomes active, favoring subject-object order over object-subject order.[9]

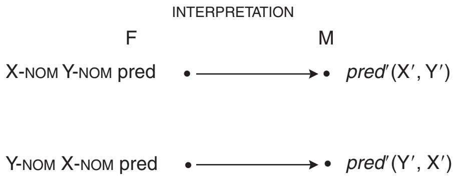
What does bidirectional optimization predict for sentences like (9)? In Strong OT the two surface forms that correspond to winners of different production optimizations are evaluated in comprehension optimization. As illustrated in the diagram below, both forms ('X-NOM Y-ACC pred' and 'Y-ACC X-NOM pred') are interpreted as having the same underlying structure, a structure corresponding to the original input to production. Any alternative interpretation, for example a candidate which interprets an accusative NP as an agent, would violate higher ranked faithfulness constraints on case interpretation and case markedness constraints, and hence is eliminated from the competition.

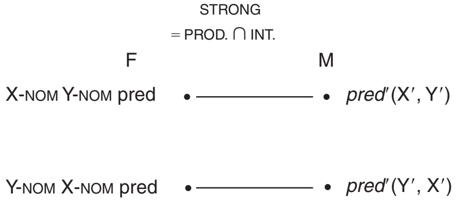




However, applying optimization in both directions produces rather surprising results for sentences with arguments that are identically case marked. For such cases, high-ranking faithfulness constraints on case interpretation and markedness constraints penalizing marked grammatical function/case associations are inapplicable (hence inactive) and low-ranking constraints that prefer canonical word order become decisive. The result is the subject-object interpretation of potentially ambiguous strings. The marked object-subject interpretation is eliminated not because it violates high-ranking faithfulness constraints, but because it violates low-ranking alignment constraints. We can illustrate this graphically as follows:







Lee (2001a) thus argues that if we define grammaticality in terms of bidirectional optimization, word order freezing within particular languages can be accounted for as an 'emergence of the unmarked' (McCarthy and Prince, 1994) in interpretation-based optimization, based on the same set of constraints that characterize cross-linguistic variation in case patterns and word order.

In sum, Strong OT offfers a unified approach to the problems of ineffability, uninterpretability, total blocking and freezing. However, Strong OT does not help with ambiguity and optionality. Since the set of Strong OT meaning–form pairs is a subset of those provided by naive interpretation for a given constraint set, Strong OT deals with ambiguity as badly as naive interpretation does. And since the set of Strong OT meaning–form pairs is a subset of those provided by naive production, it does not account for optionality either.

A related problem of Strong OT, pointed out by Blutner (2000), is that the blocking effect is so strict. For example, Strong OT predicts that "cause to die", since it is blocked by the lexicalized "kill", should be uninterpretable. But in fact it is only partially blocked, and comes to have an application in situations where "kill" would be deemed inappropriate. We now turn to Blutner's proposed solution to this problem.

4 Weak bidirectional optimization

Blutner's *weak* notion of optimality, which we refer to simply as Weak OT, is an iterated variant of Strong OT that produces partial blocking instead of strict blocking. In Weak OT, suboptimal candidates in a strong bidirectional competition can become winners in a second or later round of optimization. As we will see, in Weak OT, everyone is a winner.

Strong OT picks out a set of form–meaning pairs such that none of them is beaten by any form–meaning pair in GEN in either direction of optimization. Weak OT picks out a larger set of form–meaning pairs such that no member of that set beats any other member of the set in either direction of optimization. Thus some of the Weak OT optimal pairs may be beaten by other pairs in GEN. One may say that some Weakly optimal pairs are suboptimal. Crucially, these suboptimal optimal pairs can only be beaten by form–meaning pairs that are themselves blocked. For example, the pair ⟨"cutter", *non-canonical cutting implement*⟩ could be weakly optimal, even though it might be beaten by the pair ⟨"cutter", *knife*⟩ in a full competition amongst pairs in GEN. But this is only possible if the latter pair is itself blocked, for example beaten by the pair ⟨"knife", *knife*⟩.

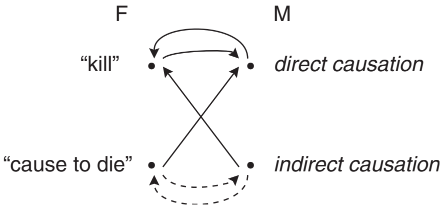
The formal definition of optimality in Weak OT runs along similar lines to the Strong OT definition, but is recursive:

(11) ⟨*f, m*⟩ is Weak OT optimal iff:

a. ⟨*f, m*⟩ ∈ GEN,

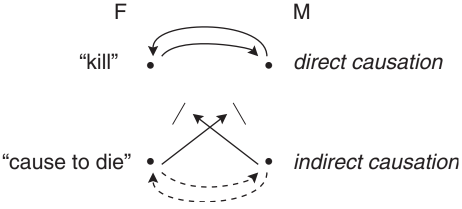
b. there is no Weak OT optimal $\langle f', m\rangle \in$ GEN such that $\langle f', m\rangle > \langle f, m\rangle$, and
c. there is no Weak OT optimal $\langle f, m'\rangle \in$ GEN such that $\langle f, m'\rangle > \langle f, m\rangle$.

The application of Weak OT, described formally by Blutner (2000), Blutner and Jäger (1999) and Jäger and Blutner (2000), can be thought of as involving repeated pruning and grafting of links between forms and meanings. We illustrate the Weak OT pruning and grafting cycle using the example of lexical and periphrastic causatives "kill"/"cause to die" which we assume are matched on the meaning side by two possible interpretations, direct causation (canonical killing) and indirect causation (non-canonical killing). The following three diagrams illustrate three phases of weak optimization. In the first diagram, all the unidirectionally optimal links are shown. In addition to the optimal links, two links are shown with dashed lines. Both of these links are unidirectionally suboptimal at this stage, beaten by other candidates:



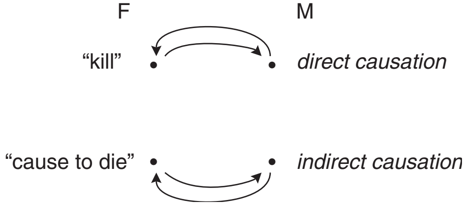
Phase 1. Naive interpretation and production

In phase 2 of Weak optimization, two unidirectionally optimal links are blocked, leaving a single bidirectionally optimal link, that between the form "kill" and the meaning corresponding to direct causation:



Phase 2. Pruning

Now we graft the originally suboptimal links between "cause to die" and the indirect causation meaning back into the picture, since the candidates which originally beat them have been removed by blocking. This gives us two bidirectionally optimal links. In the resulting happy picture, all the candidate meanings are uniquely expressible and all the candidate forms are uniquely interpretable:



Phase 3. Grafting

Blutner (2000) argues that Weak OT captures the essence of the pragmatic generalization that "unmarked forms tend to be used for unmarked situations and marked forms for marked situations" (Horn, 1984, p. 26; see also Levinson, 2000, p. 136). The concept also seems useful for deriving various alignment scales that are widely used in OT syntax work (e.g., Aissen, 1999), suggesting an interesting connection to (psychologically inspired) prototype theory. But there is a dark side to Weak OT.

First, note that Weak OT does not help with ambiguity and optionality.[10] Weak bidirectionality would predict (i) that for a form *f*, only one meaning is available if one of the meanings in pairs ⟨*f*, m_1⟩ and ⟨*f*, m_2⟩ incurs a more serious constraint violation, and (ii) that of two forms that are semantically equivalent, only one form is grammatical if one of the forms in ⟨f_1, *m*⟩ and ⟨f_2, *m*⟩ involves a more serious constraint violation. The grafting stage of Weak OT can add links to make an ineffable meaning expressible, or to give meaning to an uninterpretable form. But it cannot add new ways to express a meaning that is already expressible, or add meanings to a form that is already interpretable. So we are stuck with just the same ability to deal with ambiguity and optionality that we had in Strong OT, that is, probably not enough.

Besides this problem of undergeneration, Weak OT suffers from a more serious problem of overgeneration. Specifically, the process of adding extra links will eventually provide links for every form (if there are at least as many forms as meanings), or every meaning (if there are at least as many meanings as forms). This poses an empirical problem for uninterpretability and ineffability, and indeed also for the blocking phenomena which Weak OT was designed to account for.

The problem of overgeneration becomes intuitively clear when we apply weak bidirectionality to cases involving a fair number of form and meaning alternatives. A good example is the case pattern and relatively free word order in Korean, modeled within OT by Lee (2001a and 2003).

In canonical transitives, the case pattern in Korean is nominative-accusative, as seen in the examples in (9) above. The order of nominal arguments of the verb is relatively flexible, except for a strong verb-final restriction. However, as mentioned in Section 2, word order in this language is not random. Rather, the varied word orders are motivated by discourse and semantic factors.

Lee (2001a and 2003) models the case pattern and word order variation in Korean, assuming competing sets of case markedness constraints and alignment constraints. For our purpose here, it suffices to consider the following five constraints, ranked in the order shown in (12):

(12) a. *SUBJ/ACC: Subject is not in the accusative case.
b. HEAD-R: Head aligns right in its projection (e.g., VP) (Grimshaw, 1997).
c. SO: Subject precedes object (Choi, 1999).
d. *SUBJ^{new}: Subject is not discourse-new information.
e. *OBJ^{given}: Object is not given information.

We now consider how a plausible set of forms and meanings, shown in (13), are evaluated with respect to the constraints in (12) in Weak OT. The six forms differ in argument-case association and the surface order of the head and argument NPs; the five meanings differ in argument-function association and the givenness of arguments:[11]

(13)

forms	meanings
f_1: X-NOM Y-ACC V	m_1: *pred′*(X′, Y′), X = S^{given}, Y = O^{given}
f_2: X-ACC Y-NOM V	m_2: *pred′*(X′, Y′), X = O^{given}, Y = S^{given}
f_3: X-ACC Y-ACC V	m_3: *pred′*(X′, Y′), *X* = *O*^{new}, Y = S^{given}
f_4: X-NOM V Y-ACC	m_4: *pred′*(X′, Y′), X = *S*^{new}, Y = O^{new}
f_5: X-ACC V Y-NOM	m_5: *pred′*(X′, Y′), X = *O*^{new}, Y = S^{new}
f_6: X-ACC V Y-ACC	

Of 30 possible pairs of forms and meanings in (13), we will consider just the evaluation of 12 pairs in the tableaux that follow. They are shown in Tableau 1. Candidates labeled with the same alphabetical letter share the same meaning and differ only in positioning of the verb. (● indicates a candidate blocked by another candidate with the same form and ○, a candidate blocked by another candidate with the same meaning; ☞ marks a bidirectionally optimal form–meaning pair.) Due to bidirectional optimization, the evaluation procedure is somewhat different from standard OT: checking whether a form–meaning pair is optimal requires simultaneous evaluations of form alternatives and meaning alternatives. Tableau 1 corresponds loosely to Phase 1 of the treatment of "kill"/"cause to die" above, showing which candidates are superior in both comprehension and production. Candidate (a), with the given nominative subject and the given accusative object, emerges immediately as a bidirectionally optimal form–meaning pair.

(14) Tableau 1. First round of optimization (Weak OT)

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new},	*OBJ^{given}
☞ a. S/NOM$^{given}_1$ O/ACC$^{given}_2$ V ($\langle f_1, m_1\rangle$)					*
b. O/ACC$^{given}_1$ S/NOM$^{given}_2$ V ($\langle f_2, m_2\rangle$)			*		*
c. O/ACC$^{given}_2$ S/NOM$^{given}_1$ V ($\langle f_2, m_1\rangle$)			*		*
d. S/NOM$^{new}_2$ O/ACC$^{given}_1$ V ● ($\langle f_1, m_3\rangle$)				*	*
e. S/ACC$^{new}_1$ O/ACC$^{new}_2$ V ($\langle f_3, m_4\rangle$)	*			*	
f. S/ACC$^{new}_2$ O/ACC$^{new}_1$ V ($\langle f_3, m_5\rangle$)	*			*	
a′. S/NOM$^{given}_1$ V O/ACC$^{given}_2$ ○($\langle f_4, m_1\rangle$)		*			*
b′. O/ACC$^{given}_1$ V S/NOM$^{given}_2$ ($\langle f_5, m_2\rangle$)		*	*		*
c′. O/ACC$^{given}_2$ V S/NOM$^{given}_1$ ($\langle f_5, m_1\rangle$)		*	*		*
d′. S/NOM$^{new}_2$ V O/ACC$^{given}_1$ ($\langle f_4, m_3\rangle$)		*		*	*
e′. S/ACC$^{new}_1$ V O/ACC$^{new}_2$ ($\langle f_6, m_4\rangle$)	*	*		*	
f′. S/ACC$^{new}_2$ V O/ACC$^{new}_1$ ($\langle f_6, m_5\rangle$)	*	*		*	

In the Strong bidirectional model, we would already be finished. But in Weak OT, we have to consider the next best candidates in competitions that do not involve links blocked by the bidirectionally optimal candidate (candidate (a)). Recall that under weak bidirectionality the structures that compete in production based optimization are constrained by the outcomes of interpretation-based optimization and vice versa. Hence candidates (d) and (a′), which lose out to candidate (a) in either direction, are not contained in the candidate set for further optimization procedures. Furthermore, we remove the winning candidate (a) from the tableau: it should not be compared directly with any of the remaining candidate pairs, since it has neither the same form nor the same meaning as any of them. Hence, we arrive at the tableau in (15). As can be seen, candidate (b), in which the given accusative object precedes the given nominative subject, is selected as a winner, though it could not win under Strong OT. This is a desirable result, as this candidate, even if it violates the canonical word order requirement, is clearly a grammatical option for Korean.

(15) Tableau 2. Second round of optimization (Weak OT)

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new} ,	*OBJ^{given}
☞ b. O/ACC$^{given}_{1}$ S/NOM$^{given}_{2}$ V ($\langle f_2, m_2\rangle$)			*		*
c. O/ACC$^{given}_{2}$ S/NOM$^{given}_{1}$ V ($\langle f_2, m_1\rangle$) ●			*		*
e. S/ACC$^{new}_{1}$ O/ACC$^{new}_{2}$ V ($\langle f_3, m_4\rangle$)	*			*	
f. S/ACC$^{new}_{2}$ O/ACC$^{new}_{1}$ V ($\langle f_3, m_5\rangle$)	*			*	
b′. O/ACC$^{given}_{1}$ V S/NOM$^{given}_{2}$ ($\langle f_5, m_2\rangle$) ○		*	*		*
c′. O/ACC$^{given}_{2}$ V S/NOM$^{given}_{1}$ ($\langle f_5, m_1\rangle$) ○		*	*		*
d′. S/NOM$^{new}_{2}$ V O/ACC$^{given}_{1}$ ($\langle f_4, m_3\rangle$)		*		*	*
e′. S/ACC$^{new}_{1}$ V O/ACC$^{new}_{2}$ ($\langle f_6, m_4\rangle$)	*	*		*	
f′. S/ACC$^{new}_{2}$ V O/ACC$^{new}_{1}$ ($\langle f_6, m_5\rangle$)	*	*		*	

However, the process of recursion continues, and produces unintuitive consequences. Tableaux 3, 4 and 5 below show what happens when we consider next best candidates, even though we already found the best two. What we find is that there are many candidates generated by the Weak OT system that are not grammatical in the language modeled: none of the winners in Tableaux 3, 4 and 5 are acceptable. This shows that the present form of Weak OT is highly problematic as a model of synchronic linguistic competence.

(16) Tableau 3. Third round of optimization (Weak OT)

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new} ,	*OBJ^{given}
e. S/ACC$^{new}_{1}$ O/ACC$^{new}_{2}$ V ($\langle f_3, m_4\rangle$)	*			*	
f. S/ACC$^{new}_{2}$ O/ACC$^{new}_{1}$ V ($\langle f_3, m_5\rangle$)	*			*	

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new},	*OBJ^{given}
☞ d′. S/NOM$^{new}_{2}$ V O/ACC$^{given}_{1}$ (⟨f_4, m$_3$⟩)		*		*	*
e′. S/ACC$^{new}_{1}$ V O/ACC$^{new}_{2}$ (⟨f_6, m$_4$⟩)	*	*		*	
f′. S/ACC$^{new}_{2}$ V O/ACC$^{new}_{1}$ (⟨f_6, m_5⟩)	*	*		*	

(17) Tableau 4. Fourth round of optimization (Weak OT)

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new},	*OBJ^{given}
☞ e. S/ACC$^{new}_{1}$ O/ACC$^{new}_{2}$ V (⟨f_3, m_4⟩)	*			*	
☞ f. S/ACC$^{new}_{2}$ O/ACC$^{new}_{1}$ V (⟨f_3, m_5⟩)	*			*	
e′. S/ACC$^{new}_{1}$ V O/ACC$^{new}_{2}$ ○ (⟨f_6, m_4⟩)	*	*		*	
f′. S/ACC$^{new}_{2}$ V O/ACC$^{new}_{1}$ ○ (⟨f_6, m_5⟩)	*	*		*	

(18) Tableau 5. Fifth round of optimization (Weak OT)

	*SUBJ/ACC	HEAD-R	SO	*SUBJ^{new},	*OBJ^{given}
☞ g′. … (V O S?)					
h′. …					

The problem of overgeneration just mentioned obviously affects accounts of phenomena other than Korean word order freezing. Before closing this section, we discuss its significance for ineffability.

There have been several proposals within standard OT to deal with cases of ineffability. Among these proposals are reference to null parses

(Prince and Smolensky, 1993), the assumption of LF-unfaithful candidates (Legendre, Smolensky and Wilson, 1998), and the postulation of the lexical control component that is imposed on the optimal candidates computed by EVAL (Orgun and Sprouse, 1999). The addition of the control component may be called for independently to deal with cases of ineffability which arise from the absence of certain lexical items,[12] whereas the former two amendments of standard OT have been criticized as highly problematic from linguistic and learnability points of view (e.g., Kuhn, 2001b).

Smolensky, in unpublished work (Smolensky, 1998), has proposed a solution to language-particular ineffability, based on bidirectional optimization. What we will show is that even though a bidirectional approach may be merited, Weak OT does not fit the bill.

Recall the discussion of multiple *Wh*-questions in Italian, illustrated in (1): while English has single clause multiple *Wh*-questions, Italian does not. This is because in Italian, a markedness constraint that is violated by multiple *Wh*-questions in a single clause (Legendre, Smolensky and Wilson call this *ABSORB) is ranked higher than a faithfulness constraint to the *Wh*-feature in the input (PARSE(WH)). According to the analysis of Legendre, Smolensky and Wilson (1998), English resolves the conflict at the cost of violating the markedness constraint. Another option for resolving the conflict is by adjoining both *Wh*-phrases in [Spec, CP], as Bulgarian does. However, this option is unavailable in Italian and violates another markedness constraint *ADJOIN, which also dominates PARSE(WH) in Italian.[13]

Let us now look at what is predicted by Weak bidirectional optimization. Here, we will just go through a simplified analysis to illustrate the general effects of Weak OT; the table in (19) shows some sample forms and meanings that are relevant to our discussion:

(19)

forms	meanings
f_1: *who ate what*	m_1: $?xy ate(x, y)$
f_2: *who ate something*	m_2: $?x\exists y ate(x, y)$
f_3: *who what ate*	m_3: $?x ate(x, y)$
f_4: *who ate*	m_4: $?x ate(x, y), y = familiar$

The bidirectional competition for possible form–meaning pairs is shown in Tableau 6. With the ranking in (20), candidate (b2) is correctly predicted to be the winner in Italian; candidate (b3) is selected also as the winner, but the small set of constraints we use here does not differentiate it from (b2):

(20) Ranking for Italian: *[wh wh], *ADJOIN ≫ PARSE(WH) ≫ MARK-FAM[14] ≫ PARSE

(21) Tableau 6. Multiple *Wh*-questions in Italian (Weak OT)

	*[WH WH], *ADJOIN	PARSE(WH)	FAM-DEF	PARSE
a1. ⟨*who ate what, ?xyate*(*x*,*y*)⟩	*			
b1. ⟨*who ate something, ?xyate*(*x*, *y*?)⟩ ●		*		
c1. ⟨*who what ate, ?xyate*(*x*, *y*)⟩	*			
d1. ⟨*who ate, ?xyate*(*x*, *y*)⟩		*		*
a2. ⟨*who ate what, ?x∃yate*(*x*, *y*)⟩ ○	*			
☞ b2. ⟨*who ate something, ?x∃yate*(*x*, *y*)⟩				
c2. ⟨*who what ate, ?x∃yate*(*x*, *y*)⟩ ○	*			
d2. ⟨*who ate, ?x∃yate*(*x*, *y*)⟩ ○				*
a3. ⟨*who ate what, ?xate*(*x*, *y*)⟩ ○	*			
☞ b3. ⟨*who ate something, ?xate*(*x*, *y*)⟩				
c3. ⟨*who what ate, ?xate*(*x*, *y*)⟩ ○		*		
d3. ⟨*who ate, ?xate*(*x*, *y*)⟩ ○				*
a4. ⟨*who ate what, ?xate*(*x*, *y*), *y* = *familiar*⟩	*		*	
b4. ⟨*who ate something, ?xate*(*x*, *y*), *y* = *familiar*⟩ ●			*	
c4. ⟨*who what ate, ?xate*(*x*, *y*), *y* = *familiar*⟩	*		*	
d4. ⟨*who ate, ?xate*(*x*, *y*), *y* = *familiar*⟩			*	*

After the bidirectionally optimal candidates (b2) and (b3) have been removed from the candidate sets, candidate (d4), which could not win in the first round of optimization, becomes the winner:

(22) Tableau 7. Multiple *Wh*-questions in Italian (Weak OT)

	*[wh wh], *Adjoin	Parse(wh)	Def-Fam	Parse
a1. ⟨*who ate what, ?xyate*(*x,y*)⟩	*			
c1. ⟨*who what ate, ?xyate*(*x, y*)⟩	*			
d1. ⟨*who ate, ?xyate*(*x, y*)⟩ ●		*		*
a4. ⟨*who ate what, ?xyate*(*x, y*), *y* = *familiar*⟩ ○	*		*	
c4. ⟨*who what ate, ?xyate*(*x, y*), *y* = *familiar*⟩ ○		*	*	
☞ d4. ⟨*who ate, ?xyate*(*x, y*), *y* = *familiar*⟩ ○			*	*

The third competition certainly does not give us the correct result for Italian. As Tableau 8 shows, it predicts that multiple *Wh*-questions (f_1 and f_3) are the optimal expression for the multiple *Wh*-input ?*xyate*(*x, y*), and ?*xyate*(*x, y*) is the optimal meaning for the relevant multiple *Wh*-question. For Italian, these are unwelcome predictions:

(23) Tableau 8. Multiple *Wh*-questions in Italian (Weak OT)

	*[wh wh], *Adjoin	Parse(wh)	Def-Fam	Parse
☞ a1. ⟨*who ate what, ?xyate*(*x,y*)⟩	*			
☞ c1. ⟨*who what ate, ?xyate*(*x, y*)⟩	*			

It is not hard to see that ineffability is predicted by Weak OT only if all possible realizations for an input representation are optimal for some other meanings. As Kuhn (2001b) points out, however, this does not give us the correct result for Italian, because all strings, including *Che ha mangiato che cuesta*, are predicted to be grammatical for some other meanings.

Furthermore, Weak OT does not predict any difference between Italian and English: candidates (a1) and (b2) are predicted to be grammatical in both Italian and English under different rankings.

Although we will not provide detailed analyses, it should be obvious that these same overgeneration problems would affect the Weak OT analysis of total blocking. While in the first phase of optimization the successful Strong OT predictions appear to be reproduced, in latter stages peculiar new form–meaning pairs will emerge as winners. Provided the set of candidate meanings is large, Weak OT never predicts total blocking: all blocking is partial. So "writed", for example, would presumably be the correct expression of some meaning in Strong OT.

There remains one chink of light for Weak OT: word order freezing is still predicted, as in Strong OT, and so, for example, a Korean double nominative construction is predicted to have only a subject-object interpretation. Consider in the abstract the two forms X-NOM Y-NOM pred and Y-NOM X-NOM pred: both of these forms will be paired with meanings in the first phase of Weak optimization, so neither will enter into later competitions, and neither will become associated with incorrect argument mappings.

5 Interpretability as a constraint on production

In this and the following section we consider asymmetric models of bidirectional OT in which interpretation and production optimizations are understood to be applied in sequence, such that the first optimization affects the candidate set for the second.

Wilson (2001) discusses a model in which interpretation precedes production.[15] We refer to this as Asymmetric OT (I(nterpretation)P(roduction)).[16] In more detail, the idea of Asymmetric OT (IP) is as follows: (i) Interpretation: Given any form–meaning pair ⟨*f, m*⟩, find the most harmonic semantic interpretation of *f*. (ii) Production: Given input meaning *m*, take as candidate outputs the set of forms *f* such that ⟨*f, m*⟩ is optimal in Stage 1, and perform standard OT production optimization with this restricted candidate set. Note that the set of optimal form–meaning pairs in production is a subset of the optimal form–meaning pairs in interpretation. The set of meanings which are in some optimal pair is the same in interpretation and production, although the number of forms would, for constraint sets which are of interest, be smaller in production than in comprehension. It is the reduced set of forms in production, those which result from the two-stage process, which are to be considered grammatical, even though there are others which are interpretable.

Wilson (2001) uses this version of OT to model certain cases of partial blocking. In what follows we briefly review the Asymmetric OT (IP) treatment of partial blocking involving relativized minimality (see example (6)) and referential economy in anaphor binding. An example of a referential economy effect is provided by the following contrast between the Icelandic third person pronoun *hann* and the anaphor *sig*:

(24) Referential economy in Icelandic (Maling, 1984, p. 212):

a. Haraldur$_i$ skipaδi mér aδ raka *hann$_i$/sig$_i$.
Harold ordered me to shave him/ANAPHOR
'Harold ordered me to shave him.'

b. Jón$_i$ veit aδ María elskar hann$_i$/*sig$_i$.
Jón knows that María loves him/ANAPHOR
'Jon knows that Maria loves him.'

In (24a), the matrix subject *Haraldur* can grammatically bind the anaphor but not the pronoun. In (24b), in contrast, the pronoun is grammatical.

According to Wilson, contrasts like the one in (24) follow from an interaction of two constraints: the Local Antecedent constraint (25a), which is a locality requirement on anaphor binding, and the Referential Economy constraint (25b), which requires a bound element to be an anaphor:

(25) a. Local Antecedent: If a syntactic domain of type δ contains an anaphor α, then it also contains an antecedent for α.
b. Referential Economy: An argument does not have any lexical agreement feature specifications.

The ranking that Wilson assumes for partial blocking in anaphor binding is:

(26) Referential Economy >> Local Antecedent

The main effects of these constraints in anaphor binding are as follows. When a binding relation is suficiently local (e.g., as in (24a), when it crosses only the boundary of an infinitival clause), an anaphor need not be bound within the infinitival clause that contains it. In such a case, the anaphor, by virtue of being lexically devoid of certain agreement features,[17] is preferred to the pronoun by referential economy. But when the binding relation is non-local, as in (24b), the anaphor is excluded by Local Antecedent and the bound element must be realized as a pronoun. However, unidirectional production would predict that the non-local bound-variable interpretation is always expressed with an anaphor, since it is less marked than a pronoun in terms of referential economy.

Strong OT suffers from the same problem of strict blocking. The following tableaux will be useful for contrasting the Strong OT analysis and the Asymmetric OT (IP) treatment of the anaphora data above (to be discussed shortly) more clearly. Consider first the tableaux in (27) and (28), which illustrate interpretation optimizations based on two forms containing bound elements (an anaphor (f_1) and a pronoun (f_2)). There are two potential antecedents, one within the minimal finite clause, here labeled δ and one outside that clause. The two candidates we consider are the local binding interpretation (m_1) and the non-local binding interpretation (m_2).

For the interpretation optimization in Tableau 9, Referential Economy has no effect, since both candidates contain a bound anaphor. Thus, Local Antecedent gives us candidate (a) as the winner:

(27) Tableau 9. Interpretation I (Strong OT)

Input: [A [δ B … anaphor]] (f_1)	Referential Economy	Local Antecedent
☞ a. [A $_i$[δ B$_j$ … anaphor$_j$]] ($\langle f_1, m_1\rangle$)		
b. [A $_i$[δ B$_j$ … anaphor$_i$]] ($\langle f_1, m_2\rangle$)		*

In the interpretation optimization with the string containing a pronoun as the input, both candidates have the same constraint profile for Referential Economy and Local Antecedent, so both are selected as winners:

(28) Tableau 10. Interpretation II (Strong OT)

Input: [A [δ B … pronoun]] (f_2)	Referential Economy	Local Antecedent
☞ a. [A $_i$[δ B$_{\mathrm{j}}$ … pronoun$_{\mathrm{j}}$]] ($\langle f_2, m_1\rangle$)	*	
☞ b. [A $_i$[δ B$_j$ … pronoun$_i$]] ($\langle f_2, m_2\rangle$)	*	

In production optimizations based on m_1 and m_2, on the other hand, due to the higher ranking constraint Referential Economy, the same candidate (a) wins for both inputs:

(29) Tableau 11. Production I (Strong OT)

Input: local binding (m_1)	Referential Economy	Local Antecedent
☞ a. [A $_i$[δ B$_j$ … anaphor$_j$]] ($\langle f_1, m_1\rangle$)		
b. [A $_i$[δ B$_j$ … pronoun$_{\mathrm{i}}$]] ($\langle f_2, m_1\rangle$)	*	

(30) Tableau 12. Production II (Strong OT)

Input: non-local binding (m_2)	Referential Economy	Local Antecedent
☞ a. [A $_i$[δ B$_j$… anaphor$_i$]] (⟨f_1, m_2⟩)		*
b. [A $_i$[δ B$_j$… pronoun$_i$]] (⟨f_2, m_2⟩)	*	

Thus Strong OT produces only one bidirectionally optimal form–meaning pair, that is, ⟨f_1, m_1⟩, failing to predict partial blocking.

Wilson (2001) offers an Asymmetric OT (IP) account of these facts that overcomes these problems. Crucially, in Wilson's model, interpretation optimization applies first to limit the candidate set for the second, production optimization. To see how the analysis works, compare the tableaux in (29) and (30) with the ones in (31) and (32) below, which correspond to the second stage of optimization in Asymmetric OT (IP).[18] As we noted above, in the Strong OT model, the results of optimization under one direction does not affect which candidates compete under the other direction because the candidate set of both directions of optimization is defined independently. Consequently, all the four form–meaning pairs in the above interpretation tableaux compete under the production optimization also. But in Asymmetric OT (IP), only winning candidates in interpretation enter into the production optimization.

For the anaphora data under discussion here, the consequence of this is as follows: since m_2 loses in the interpretation tableau with input f_1 (Tableau 9), the production competition with m_2 as input no longer includes the candidate f_1. That is, the original production tableau which took m_2 as input (Tableau 12) must be replaced by Tableau 14, which does not include candidate (a). As a result, candidate (b) wins trivially, and m_2 is predicted to be realized as f_2. Meanwhile, the production tableau for meaning m_1 (Tableau 11) is unaffected, so m_1 is still realized as f_1:

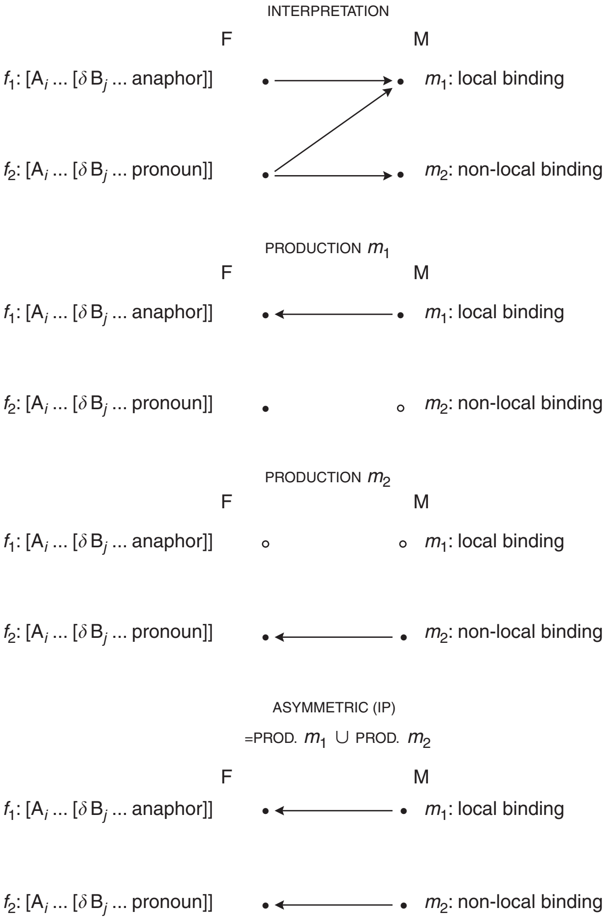
(31) Tableau 13. Production I (Asymmetric OT (IP))

Input: local binding (m_1)	Referential Economy	Local Antecedent
☞ a. [A $_i$[δ B$_j$… anaphor]] (⟨f_1, m_1⟩)		
b. [A $_i$[δ B$_j$… pronoun$_i$]] (⟨f_2, m_1⟩)	*	

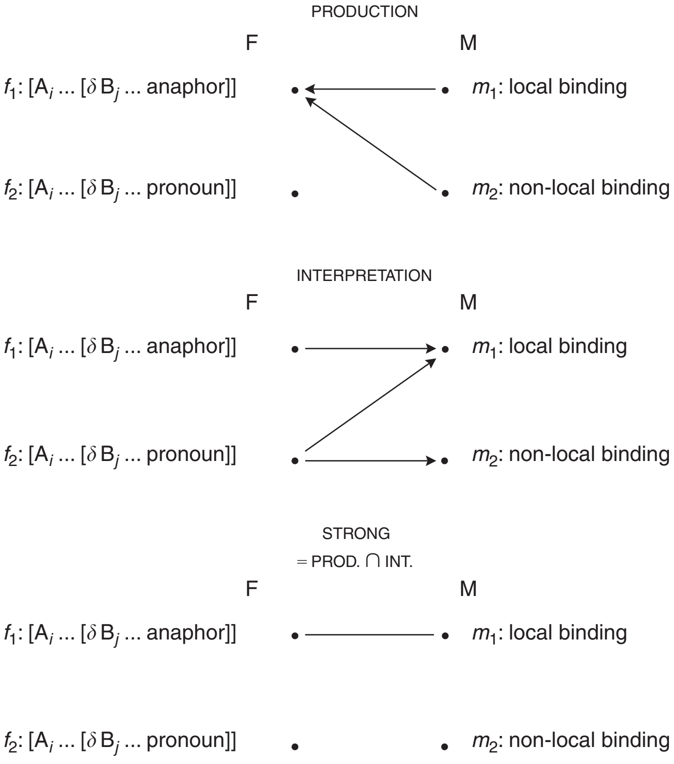
(32) Tableau 14. Production II (Asymmetric OT (IP))

Input: non-local binding (m_2)	REFERENTIAL ECONOMY	LOCAL ANTECEDENT
☞ b. [A $_i$[δ B$_j$ … pronoun$_i$]] (⟨f_2, m_2⟩)	*	

The process Wilson describes is pictured in the following diagram, where candidates are marked using "o" for those competitions where they are not participants:


INTERPRETATION
F
M
f_1: [A$_i$ ... [δ B$_j$ ... anaphor]]
m_1: local binding
f_2: [A$_i$ ... [δ B$_j$ ... pronoun]]
m_2: non-local binding
PRODUCTION m_1
F
M
f_1: [A$_i$ ... [δ B$_j$ ... anaphor]]
m_1: local binding
f_2: [A$_i$ ... [δ B$_j$ ... pronoun]]
m_2: non-local binding
PRODUCTION m_2
F
M
f_1: [A$_i$ ... [δ B$_j$ ... anaphor]]
m_1: local binding
f_2: [A$_i$ ... [δ B$_j$ ... pronoun]]
m_2: non-local binding
ASYMMETRIC (IP)
=PROD. m_1 ∪ PROD. m_2
F
M
f_1: [A$_i$ ... [δ B$_j$ ... anaphor]]
m_1: local binding
f_2: [A$_i$ ... [δ B$_j$ ... pronoun]]
m_2: non-local binding

We may compare Wilson's successful account of referential economy with the results that would be obtained in Blutner's models. Whereas Weak OT, which deals quite effectively with partial blocking, would successfully predict the Icelandic data, Strong OT would be less successful. As the following diagram shows, under the constraints assumed, Strong OT incorrectly predicts that Icelandic pronouns are uninterpretable in the given configuration, and that there is no way of expressing non-local binding:


PRODUCTION
F
M
f_1: [A_i ... [δ B_j ... anaphor]]
m_1: local binding
f_2: [A_i ... [δ B_j ... pronoun]]
m_2: non-local binding
INTERPRETATION
F
M
f_1: [A_i ... [δ B_j ... anaphor]]
m_1: local binding
f_2: [A_i ... [δ B_j ... pronoun]]
m_2: non-local binding
STRONG
= PROD. ∩ INT.
F
M
f_1: [A_i ... [δ B_j ... anaphor]]
m_1: local binding
f_2: [A_i ... [δ B_j ... pronoun]]
m_2: non-local binding

So far we have looked at the Asymmetric OT (IP) analysis of partial blocking in anaphor binding. What of the standard cases of partial blocking we considered earlier? Can they be modeled in Asymmetric OT (IP)? It is interesting to note that all cases of partial blocking are subject to two similar kinds of constraints: one that favors a less marked form and the other that favors a less marked meaning. In the case of Icelandic anaphor binding, REFERENTIAL ECONOMY concerns formal markedness, and LOCAL ANTECEDENT concerns semantic markedness; in the example of causatives discussed in the previous section, the formal markedness constraint was a preference for short forms, and the semantic markedness constraint was a preference for the canonical mode of causation.

Yet there is an important difference between the phenomena Wilson models and the partial blocking cases considered earlier. What distinguishes Wilson's anaphora data is that the pair of a marked form and an unmarked meaning (⟨f_2, m_1⟩ in the above tableaux) and the pair of a marked form and a marked meaning (⟨f_2, m_2⟩ in the above tableaux) have the same constraint profile for the constraint favoring a less marked meaning (see Tableaux 9 and 10 above; see also Wilson (2001, pp. 496–8) for a detailed discussion). As noted above, the LOCAL ANTECEDENT constraint, preferring local binding over non-local binding, targets only an anaphor (f_1) but not a pronoun (f_2). As a result, the pairs ⟨f_2, m_1⟩ and ⟨f_2, m_2⟩ both survive in interpretation. Now when we come to realize m_1, we don't choose f_2 but instead choose f_1. In other words, in production, as illustrated in Tableaux 13 and 14, the pair ⟨f_1, m_1⟩ blocks ⟨f_2, m_1⟩, making ⟨f_2, m_2⟩ available.

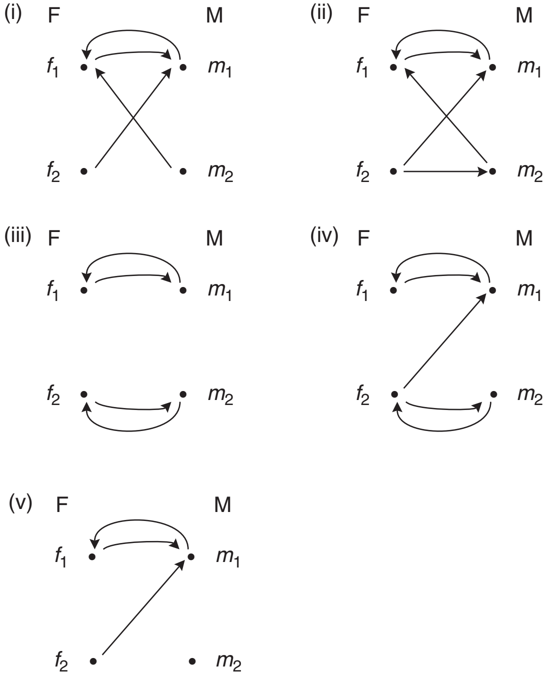
The standard cases of partial blocking differ in that the two pairs ⟨*marked form, unmarked meaning*⟩ and ⟨*marked form, marked meaning*⟩ do not have the same constraint profile. This is illustrated in (33):

(33) Tableau 15. Interpretation

Input: *cause to die*	ECONOMY	CANON
☞ a. ⟨*cause to die, direct causation*⟩	*	
b. ⟨*cause to die, indirect causation*⟩	*	*

Asymmetric OT (IP) fails to predict the full "division of pragmatic labor" whereby more marked forms are associated with more marked meanings. The constraints above yield a preferred interpretation of "cause to die" as involving canonical direct causation. Therefore, in the production competition with indirectly caused death as input meaning, "cause to die" is not even amongst the candidate outputs, and cannot be the winner. Presumably, the winner would be some even more periphrastic alternative such as "indirectly cause to die".

We can see the difference between the two cases, and how they are treated, graphically. Diagrams (i)–(v), below, show both production and interpretation relations. The first two diagrams represent direct applications of naive back-and-forth OT. The first illustrates standard partial blocking cases yielding marked meanings for marked forms such as "cutter" and "cause to die".



The second diagram represents the situation Wilson describes for Icelandic anaphora. The only difference is an extra arrow from the marked form to the marked meaning in the second diagram.

Diagram (iii) shows the results of applying Weak OT to either the situation in (i) or that in (ii): the marked form becomes uniquely associated with the marked meaning in both directions of optimization, while the unmarked form and unmarked meaning continue to be a bidirectionally optimal pair as they were in the original cases. Asymmetric OT (IP) does not achieve the harmonious situation depicted in (iii) for either of the situations given by (i) and (ii). What it does achieve is represented in (iv) and (v). Diagram (iv) shows the results of applying Asymmetric OT (IP) to the Icelandic anaphora case in (ii). Here we see that the division of labor depicted in (iii) is almost achieved, except that there remains the possibility of interpreting the marked form as the unmarked meaning. This is a result of the fact that Wilson's proposal does not innovate above naive back-and-forth OT as regards interpretation. When Asymmetric OT (IP) is applied to the classic "cause to die" situation in (i), what results is (v). Wilson's system does not succeed in creating any link between the marked form and the marked meaning, so we can see that it does not provide a very general model of partial blocking. In these cases we might better describe what it does as "almost blocking".

Asymmetric OT (IP) has an interesting range of strengths and weaknesses. We have just seen that it produces mixed results with respect to partial blocking. It does not help with ambiguity and optionality, since it does not provide new meanings for a form already contained in the set of winners in interpretation, or provide new ways to express a meaning that is already in the set of winners in interpretation. It also does not predict uninterpretability, since interpretation is naive. On the other hand, Wilson's system can help with total blocking and freezing. Consider, for example, the two Korean double nominative forms X-NOM Y-NOM pred and Y-NOM X-NOM pred: both of these forms will be paired with the subject-object interpretation in the first, interpretation stage of optimization. So the pairs of these forms and the object-subject interpretation will not be included in the legitimate candidate set for the second, production optimization, and we derive the effect of freezing. Ineffability is predicted in some cases. Suppose a meaning is highly marked, such that no form is interpreted as having that meaning. In this case Asymmetric OT (IP) predicts that with this form as input, there will be no output (since there will be no candidates at all in the second stage of the production optimization). But it is not obvious whether this is sufficient to account, for example, for the ineffability of multiple *Wh*-questions in Italian.

6 Reproducibility as a constraint on interpretation

Zeevat (2000), like Wilson (2001), suggests using entirely different architectures for production and interpretation. What is striking is that Zeevat and Wilson choose precisely opposite architectures. Wilson keeps the standard unidirectional OT model of interpretation, but restricts the candidate set for production using the results of interpretation. Zeevat keeps the standard unidirectional OT model of production, but restricts the candidate set for interpretation using the results of production.

Zeevat bases his argument for what we will term Asymmetric OT (PI) in large part on two phenomena we have been discussing in this paper, ambiguity and ineffability. As regards ambiguity, we can gloss the idea as follows: since naive OT production has no problem with ambiguity, we should use the production architecture as the basis of comprehension, and add further interpretational bells and whistles only as necessary.

In more detail, Zeevat's model starts by assuming that production uses a standard OT syntax set of constraints that we will term PROD. Comprehension is a more involved two-stage process involving both PROD and an additional set of constraints to select between alternative meanings: we will refer to this second set as PRAG. Zeevat's use of two distinct constraint sets for interpretation and production amounts to a significant difference from both Wilson's proposal and the other bidirectional architectures we have discussed, although Hendriks and de Hoop (2001) also advocate such a split.

The first stage of comprehension of a form $\mathcal{F}$ consists in determining the set $\mathcal{M}$ of meaning inputs which give $\mathcal{F}$ as output using the constraints PROD. The second stage consists in using a standard OT semantics form-to-meaning optimization with the form $\mathcal{F}$ as input, except that rather than using GEN to give candidate outputs, the set $\mathcal{M}$ is used.

As is the case for Wilson's model, the form–meaning relation defined for production in Zeevat's proposal is different than that for comprehension. For Zeevat, the set of form–meaning pairs in comprehension is a subset of those in production. So a first observation on the proposal is that it predicts the existence of cases of *guaranteed misinterpretation*, that is, cases where a given meaning is expressed in a way that would be understood as having an interpretation other than the original meaning. Indeed, the proposal would seem to stand or fall on the existence of such cases, since without them the grounds for introducing a radical difference between production and comprehension are weak.

Zeevat does not cite any cases of guaranteed misinterpretation: the data he gives concerns the form–meaning relationship in the abstract, not differences between the form–meaning relationship provided by the production component of his system and the form–meaning relationship given by the comprehension system. In other words, his data involves form–meaning mismatches, like ambiguity and ineffability, not comprehension–production mismatches involving guaranteed misinterpretation.

Furthermore, while Zeevat describes the constraint set in PRAG, he does not describe PROD, so it is hard to be sure what the range of cases is where he predicts a mismatch between production and comprehension. None the less, we can exemplify the type of comprehension–production mismatch Zeevat predicts. The conjunction in (34a) involves two occurrences of an expression presupposing that there was a mosquito. A natural interpretation would involve only one mosquito, in which case the discourse might be continued with (34b), but it is also possible (if strained) to continue the discourse as in (34c), a two mosquito interpretation:

(34) a. Hanjung realized that there was a mosquito and David realized that there was a mosquito.
b. The mosquito was hungry.
c. Both mosquitos were hungry.

Although we do not know what constraints are in PROD, we can speculate that (34a) might be generated in either the one mosquito or the two mosquito model. However, Zeevat postulates a constraint *ACCOMMODATE in PRAG, a constraint which would prevent accommodation of presuppositions when the presuppositions are already satisfied in the discourse context. In this case, when the interpreter arrives at the second clause of (34a), a discourse referent for a mosquito has already been established, so there is no need to accommodate an extra mosquito in order to process the presupposition of the second clause. Thus Zeevat predicts that only the one mosquito interpretation should be available. So this may be a case where Zeevat predicts guaranteed misinterpretation. A speaker wanting to express that Hanjung and David have realized that separate mosquitos exist may optimally report this as in (34a), but in this case will be understood to mean that Hanjung and David have both developed existential knowledge about the same mosquito.

As regards (34a), the data is murky, since there is a slight awkwardness to the continuation in (34c). Our point is not to use this case to attack or defend Zeevat's account, but rather to bring out more clearly the type of prediction that would provide a test for the proposed architecture. Detailed consideration of the predictions would have to wait until we know more about PROD.

As noted, ambiguity is one of the main motivations claimed for Asymmetric OT (PI): Zeevat analyzes the *Rat/Rad* (rat/wheel) problem at length. In interpretation, it is unproblematic for both the meanings *rat* and *wheel* to be selected in the first stage of comprehension (the reverse production stage), and there is no reason to expect PRAG to produce any preference between them, so ambiguity is predicted. However, there is an important class of examples for which Zeevat's system incorrectly eliminates ambiguity. The problem is that PRAG includes a constraint STRENGTH which prefers logically stronger interpretations to weaker ones, so that Zeevat's asymmetric model never predicts that one reading of an ambiguous sentence will entail another.

Consider (35a), which by virtue of a standard quantificational scope ambiguity has the two readings in (35b) and (35c):

(35) a. Every child liked one toy,
b. $\forall x \mathrm{child}(x) \rightarrow (\exists y \mathrm{toy}(y) \vee \mathrm{liked}(x, y))$
c. $\exists y \mathrm{toy}(y) \vee; (\forall x \mathrm{child}(x) \rightarrow \mathrm{liked}(x, y))$

Here (35c) entails (35b), so Asymmetric OT (PI) incorrectly predicts that only the former is available. Furthermore, note that cases in which one reading entails another are common. Apart from scope ambiguities, this situation often arises when one meaning of a polysemous word has a strictly greater extension than another, as in "finger" (all digits on a hand, or all but the thumb), "gay" (homosexual, or homosexual male), and "New York" (the city or the state containing the city). Thus Asymmetric OT (PI) would predict that "There are rats in New York" can only mean that there are rats in New York City (and hence also in the state), while "There are no rats in New York" can only mean that there are no rats in the state (and hence none in the city either). We can conclude that while the architecture Zeevat proposes can successfully model ambiguity phenomena, the specific constraints he uses are problematic.[19]

Another claim of Zeevat's is that Asymmetric OT (PI) successfully handles ineffability. However, we find that this claim is not yet fully substantiated. Note that Zeevat's claim is based on interpretation. But if ineffability consists in the existence of meanings which cannot be realized in production, then Zeevat's model does not predict any ineffability, since from a production perspective, any meaning will give some winning form. So an Italian wanting to express the multiple *Wh*-question "Who ate what?" would be predicted to produce some utterance, and it is not obvious why this Italian would not imagine he or she had successfully expressed exactly what he or she intended.

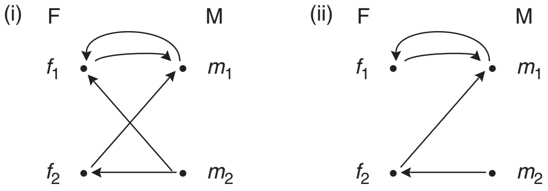
Zeevat's point, then, is more limited: in his system there may be no Italian form that would be understood as "Who ate what?" First, consider the infelicitous "Chí ha mangiato che cosa?" ("*Who ate which thing?*"). Zeevat would analyze this as uninterpretable because the PROD constraints prevent any meaning from being expressed that way. This seems reasonable. So we need to consider which string would be the output for the input ?*xy*ate(*x, y*). Zeevat supposes this to be "Chí ha mangiato qualcosa?" (literally, "*Who ate something?*"). The question is then why this string is not interpreted as ?*xy*ate(*x, y*), but instead as ?*x*∃*y*ate(*x, y*).

Given the premise that in the first stage of comprehension for the form "Chí ha mangiato qualcosa?" both these meanings are found, selection between them is left to PRAG. Zeevat assumes that the crucial constraint will be one he terms *INVENT, that will disallow a mismatch between the numbers of question variables and existentials in the meaning and the numbers of corresponding expressions in the form.

How could *INVENT achieve such careful accounting of the differences between form and meaning? One possibility would be that *INVENT incorporated many or all of the constraints in PROD, but this would call into question the basic premise that PROD and PRAG are independent constraint sets with quite different functions. Zeevat (p.c.) has suggested instead that *INVENT is defined purely on meanings, not making any reference to forms. All it is supposed to do is prefer minimal meanings, for example in the sense of requiring less structure in a DRS. On this basis, *INVENT could prevent "Chí ha mangiato qualcosa?" from being interpreted as ?*xy*ate(*x, y*), but only if there was some well-defined sense in which this meaning was less minimal than the alternative ?*x*∃*y*ate(*x, y*). We see no a priori reason why a meaning with two question variables should be less minimal than a meaning with one question variable and one existential variable, but this is perhaps not a major drawback of Zeevat's proposal. What is clear is that, in principle, the architecture Zeevat advocates is capable of partially accounting for cases of ineffability like multiple questions in Italian. We say "partially" because, as pointed out above, Asymmetric OT (PI) fails to account for why speakers do not produce forms with the intention of expressing a multiple question: it can only account for why the forms they produce are misunderstood.

We have looked at the Asymmetric OT (PI) treatment of ambiguity and ineffability: what of uninterpretability, optionality, blocking and freezing? As with all the other accounts we have considered, Zeevat's proposal has both strengths and weaknesses.

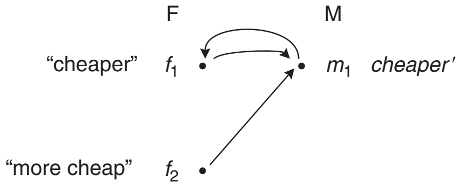
Regarding optionality, Asymmetric OT (PI) introduces no new insights above naive production OT: typically, there will be a single winning form. Also, with regard to freezing, Zeevat's model does not seem to provide a solution. In the case of Korean psychological verbs, for example, there is nothing to stop production of both SOV and OSV word orders. With regard to partial blocking, and by analogy with Wilson's system, Zeevat's proposal offers at best a partial solution. In particular, it is easily verified that Asymmetric OT (PI) makes incorrect predictions for both "cause to die" type examples and cases with the same structure as found with Icelandic anaphora. On the other hand, we can easily identify the abstract structure of two cases for which Zeevat's system would successfully isolate two form–meaning pairs from each other. Diagrams (i) and (ii) show naive back-and-forth OT structures which under Asymmetric OT (PI) would yield two bidirectional links, one between f_1 and m_1, and the other between f_2 and m_2. Identifying linguistic phenomena to which these two diagrams correspond might provide further insight into the significance of Zeevat's model, but we leave this task to future research.



Let us briefly consider the option of treating partial blocking and freezing by combining Zeevat's model with Blutner's Strong or Weak optimality. For example, we might define Strong Asymmetric OT (PI) as having the form–meaning relationship defined by the intersection of Zeevat's production and comprehension mechanisms. However, we already noted that the set of form–meaning pairs in Zeevat's production model is a superset of the form–meaning pairs in his comprehension model. So taking the intersection of the two would amount to using the comprehension model for both comprehension and production. Given the philosophical position taken in Zeevat (2000), and the many arguments he gives for an asymmetry between comprehension and production, a move to Strong Asymmetric OT (PI) would amount to something of a retreat, even if the result successfully modeled freezing. Still, we think it worth noting the possibility of such a model, as one of many directions to which Zeevat's model may be extended, and one of many possibilities in the space of bidirectional OT architectures.

Where Asymmetric OT (PI) certainly does have something to offer is with respect to uninterpretability and total blocking. Regarding uninterpretability, observe that since the first stage of interpretation is identical to naive production, there will in general be many strings which are not produced for any meaning input. All these strings are uninterpretable in Asymmetric OT (PI). For example, if "Colorless green ideas sleep furiously" is the form, we would first consider the set of meanings that would generate it. If we allow Chomsky's premise that the string is meaningless, we would find no such meaning, and hence the model correctly predicts uninterpretability (due to complete absence of any candidates in the final stage of the interpretation competition).

Last, we consider total blocking. It is easy to see that Asymmetric OT (PI) can model this phenomenon, the analysis being parallel to that of uninterpretability. Consider a standard case:



When Asymmetric OT (PI) is applied in the above situation, the interpretation arrow from "more cheap" to the meaning *cheaper*' would be removed. The reason is that when interpreting "more cheap", the only candidate meanings considered are those which would be expressed as "more cheap". By assumption, the lexicalized "cheaper" is the most harmonic expression of this meaning, so we know that the meaning is not realized as "more cheap". If there are no other meanings that would be realized as "more cheap", then once again we have a case of an empty candidate set, and "more cheap" becomes uninterpretable, effectively blocked by "cheaper".

7 Conclusions

We have reviewed the predictions of seven different versions of OT with respect to seven empirical phenomena. Our main conclusions are summarized in the following table:

Approach	Ambiguity	Optionality	Ineffability	Uninterpretability	Total blocking	Partial blocking	Freezing
Naive production	√	×	×	√	×	×	×
Naive interpretation	×	√	√	×	×	×	×
Back-and-forth	×	×	×	×	×	×	×
Strong	×	×	√	√	√	×	√
Weak	×	×	×	×	×	√	√
Asymmetric (IP)	×	×	√	×	√	√?	√
Asymmetric (PI)	√	×	√?	√	√	×	×

In interpreting the table, several caveats should be born in mind. First, we could have chosen a different set of phenomena to consider. Second, there is no interesting sense in which the seven phenomena we focused on are of equal significance. Third, some may even doubt whether certain of the phenomena constitute real problems linguistically. For example, one might take differing stances with respect to Chomsky's view that there are syntactically well-formed strings that lack an interpretation, and perhaps even doubt the existence of uninterpretability. One might say that for any string, given enough time, we could find a situation where it was appropriate to use that string. Or one might take issue with synonymy, doubting that two different expressions ever mean exactly the same thing.

So we accept that there is room for disagreement about how significant each of the seven phenomena is. Yet we also believe that a strong argument could be made for not restricting ourselves to grammar architectures that make description of these phenomena impossible. The table above shows that bidirectional OT architectures from the literature are too restrictive: there are many patterns of relation between form and meaning which they cannot describe effectively, regardless of the particular constraints that are used and the ranking between those constraints. Even the account which (narrowly) fairs best by our criteria, Strong OT, fails to contribute to our understanding of three of the seven phenomena, ambiguity, optionality and partial blocking.

In this chapter we have not attempted to present an approach which betters existing proposals. However, there is no shortage of directions in which these existing proposals could be developed. Consider, first, partial blocking. Only one of the proposals discussed, Weak OT, deals with the classic cases of partial blocking described in Section 2. Yet Weak OT suffers from severe problems, most notably considerable overgeneration. Could a variant of Weak OT maintain the analysis of partial blocking without this leading to such great overgeneration? One possibility to consider is the variant of Weak OT discussed by Beaver (to appear). This variant system performs only one iteration of the Weak OT process, pruning once and grafting once. As a result it maintains some of the properties of Weak OT, but lacks Weak OT's "everyone's a winner" profligacy.

There are also several approaches that could be combined with the proposals discussed here so as to account for optionality and ambiguity. Partial ranking of constraints (Anttila and Fong, 2000), and stochastic ranking of constraints (Boersma and Hayes, 2001; Asudeh, 2001; Bresnan and Deo, 2001) are techniques that allow multiple winners to appear in competitions that might only produce a single winner using linear constraint ranking. Another issue that is very relevant to ambiguity and optionality is the role played by context. For example, so-called *optionality* of Korean transitive word order can also be seen as context-dependence of Korean word order: in specific discourse contexts where one argument is more prominent than the other, there may be no word order freedom at all. So it is natural to move from simple form–meaning or meaning–form optimization to optimizations that include three parameters: form, meaning and context. This is exactly what Blutner (2000) proposes, although his main use of context involves presupposition resolution rather than ambiguity resolution or what we might analogously term *optionality resolution* – the context-dependent choice of a particular form from amongst a range of possibilities.

We have shown that existing bidirectional OT systems suffer from serious problems in their treatment of form–meaning asymmetries. But our chapter is intended in a constructive spirit. We have laid out a set of issues which we hope developers of bidirectional approaches will tackle in future research.

Notes

1. For a recent discussion of the many aspects of ambiguity and why they constitute a puzzle for linguistics, see Wasow et al. (to appear).
2. An occupational therapy web-site (www.otworks.com) reports that: "OT stands for…Occupational Therapy, or Over Time Ol' Timer Original Thinkers Overly Timid Old Testament Over Taxed E.T's sibling." OT is also used to mean "Off Topic".

3. See Müller (1999) for an overview of approaches to optionality within the standard OT framework whose constraint set forms a total order.
4. English complementizer drop has been analyzed within OT by Grimshaw (1997) and Baković and Keer (2001).
5. Note that the meaning de Hoop (2001) gives to the term *unintelligibility* seems, from her examples, to be distinct from our notion of *uninterpretability*. The examples de Hoop considers involve utterances which have (only) a contradictory interpretation, whereas we consider cases in which one cannot determine any proposition expressed by the utterance.
6. We choose "last" in the diagram as an arbitrary highly unmarked adjective, at least in terms of having higher frequency than any other adjective in the British National Corpus. If this can be taken to indicate that the meaning is less marked than other adjectival meanings, then OT grammars might be expected to interpret "dolomphious" as having the same meaning as "last".
7. Some discussion of different options of combining two optimization perspectives and the general consequences for the resulting bidirectional models can be found in Kuhn (2001b).
8. See also Kuhn (2001b) and Vogel (Chapter 9).
9. In this way, unidirectional production OT can produce apparent optionality, based on different inputs. This approach to optionality, which Müller (1999) terms "the pseudo-optionality approach", predicts cases of optionality that correlate with differences in information status but does not produce multiple outputs for the same input.
10. These problems of Weak OT are also discussed by Gärtner, in Chapter 7.
11. Though information about argument–function mappings is represented as part of "meanings" in (13), we do not assume that this information is part of OT input. Rather association of the arguments in the input to a particular grammatical function results from constraint interaction.
12. Some discussion of a typology of ineffabilities can be found in Fanselow and Féry (to appear).
13. Yet another option which we consider as a candidate for realizing the multiple *Wh*-question in Italian is the ellided form "Chí ha mangiato?" ("Who has eaten?"). The elliptical form, however, would express the multiple *Wh*-question only at the cost of violating a faithfulness constraint PARSE, which requires input elements to have an overt correspondent in the output (Grimshaw and Samek-Lodocivi (1998)). Clearly, in reality, this cost is too high.
14. MARK-FAM requires that familiar objects are realized as definites. This is a counterpart to the constraint FAM-DEF in Beaver (to appear) which requires that definites should be familiar. Note that MARK-FAM can penalize indefinites, whereas FAM-DEF can only penalize definites.
15. Our understanding of Wilson's model is considerably influenced by recent unpublished work of Judith Aissen.
16. Vogel (Chapter 9) develops a bidirectional OT model in which production-based optimization is accompanied by a second step that checks the recoverability of an underlying form. We defer discussion of this model to a later occasion.
17. For example, *sig* is unmarked for gender and number, and *hann* is a masculine and singular form.
18. Wilson (2001) makes two assumptions regarding representations of inputs and candidate structures in his analysis. The first is that the input for interpretation and production is the same and only the candidate set varies. More specifically,

for both optimizations, he assumes a highly abstract input consisting of a surface string plus an abstract syntactic structure (i.e., LF) and a semantic representation. In interpretation, the morphosyntactic component of the input is held fixed across the candidate set; in production, the semantic component is fixed. Second, Wilson assumes that binding relations are specified in the input semantic representation and that in interpretation candidates may diverge from the input with respect to binding relations. Relativized minimality in interpretation and referential economy in production then are both accounted for in terms of faithfulness violations. In this discussion, we abstract away from details of representational assumptions that Wilson makes and continue to assume that for interpretation, the input is a form and the output is a meaning; and for production, the input is a meaning and the output is a form. As far as we can tell, this does not affect the overall results of Asymmetric OT (IP).

19. Observe that if Asymmetric OT (PI) can model ambiguity, one might expect by symmetry considerations that Asymmetric OT (IP) would model optionality. But here the fact that Zeevat uses two distinct constraint sets while Wilson uses only one comes into play. It is because Zeevat proposes that the set of interpretation constraints is very limited that his system can model ambiguity. By contrast, Wilson uses the full constraint set in the second phase of production, and this will typically weed out all but one candidate. An architecture like that of Asymmetric OT (IP) would model optionality provided it used only a very limited constraint set in the second stage of production, and kept the bulk of constraints for the first stage of production and for interpretation.

7
On the Optimality Theoretic Status of "Unambiguous Encoding"

Hans-Martin Gärtner

1 Introduction

In the present chapter, I am going to explore the relation between pragmatics and optimality theory (henceforth OT) in the area of disambiguation. My starting point will be an analysis of Icelandic object-shift and differential marking of (in)definite theme arguments in Tagalog. I argue that OT is able to capture the interaction of interpretive and morphosyntactic constraints involved there in a particularly insightful way. More specifically, a certain functional flavor of object-shift and argument marking, both arguably carried out for the purpose of disambiguation, comes out as the "emergence of the unmarked".[1] A direct link between this property and the OT formalism will be postulated in terms of a family of (disambiguation) constraints called "Unambiguous Encoding" (henceforth UE). This is attractive to the extent that UE could be taken to be grounded in Gricean principles like "Be Perspicuous", or "Avoid Ambiguity" (Grice, 1989, p. 27).[2]

In the second part of this chapter, I will point out some shortcomings of the UE approach. These shortcomings can be taken to indicate that the OT status of UE is epiphenomenal. Two methods of reduction are explored, which, if they can be worked out in a satisfactory way, must be preferred to the UE approach.

First, Blutner's bidirectional optimization is considered (Blutner, 2000), since it provides a very attractive approach to disambiguation, the latter resulting directly from the optimization process rather than having to be built in explicitly at the level of constraints. However, in its original form this method can be shown to fail on the specific Icelandic and Tagalog patterns discussed. A possible (partial) repair by means of a contextual constraint has the consequence of eliminating ambiguity directly, and thus of weakening the disambiguative power of this method.

Second, applying Aissen's approach to "differential object marking" in terms of "harmonic alignment" and "local conjunction" (Aissen, 1999, 2000) turns out to be a more successful reduction of UE. Unfortunately, this success seems to come at the cost of eliminating the functionalist intuitions behind the UE approach. In addition, like UE but unlike weak bidirectional OT, it suffers from the excessive power of appeal to input/output constraints.

2 The rise of "unambiguous encoding"

2.1 Icelandic object-shift

Vikner (1997, 2001), building on work by Diesing (1996), presents an OT analysis of Icelandic object-shift, an example of which is (1):

(1) a. Auk þess sýna þau alltaf viðtal við Clinton í erlendu fréttunum.
Besides show they always interviews with C. in the foreign news
'Besides, in the foreign news they always show interviews with Clinton.'
b. Auk þess sýna þau viðtal við Clinton alltaf í erlendu fréttunum.
'Besides, interviews with Clinton they always show in the foreign news.'

While (1a) shows the object *viðtal við Clinton* in its VP-internal *in situ* position, (1b) is a case of object-shift. On the assumption that "medial" adverbs of quantification like *alltaf* are located at the left edge of VP (1b) shows the object in a VP-external *shifted* position. A sketchy syntactic analysis of these facts is given in (2), where (2a) and (2b) correspond to (1a) and (1b) respectively.

(2) a. ... [$_{\text{VP}}$ alltaf [$_{\text{VP}}$ t_V [$_{\text{NP}}$ viðtal við Clinton] ...]]
b. ... [Agr$_{\text{OP}}$ [$_{\text{NP}}$ viðtal við Clinton]$_{\text{j}}$ [$_{\text{VP}}$ alltaf [$_{\text{VP}}$ t_V t_j ...]]]

As can be gathered from the translations, the examples in (1) differ in meaning. This difference is captured by the paraphrases in (3), where (3a) and (3b) correspond to (1a) and (1b) respectively.

(3) a. No foreign newscast goes without an interview with Clinton.
b. No interview with Clinton escapes being broadcast in the foreign news.

Let me call (3a) the *weak* indefinite reading and (3b) the *strong* one, in keeping with common practice. Note that, under a slight idealization (1a) *only* supports reading (3a) while (1b) *only* supports reading (3b).[3]

Vikner's account presupposes among other things the familiar analysis of adverbs of quantification in terms of a tripartite quantificational structure $Q(_{restrictor})(_{nucleus})$. On this basis the weak indefinite reading results from the indefinite adding its content to the nuclear scope, while the strong one requires that content to go into the restrictor clause. In addition, one has to adopt the "Mapping Hypothesis" (Diesing, 1996), according to which VP-internal material maps into the nuclear scope while VP-external material

("IP (above VP)") maps into the restrictor clause. The facts in (1)–(3) can then be derived by means of a constraint called SCOPING (Vikner, 2001, p. 328):[4]

(4) SCOPING

An element has the (surface) position in the clause that corresponds to its scope.

As the tableaux in (5) and (6) show, SCOPING is able to associate the weak indefinite reading with the *in-situ* form and the strong indefinite reading with the shifted form:

(5)

	input: weak indefinite object	SCOPING
☞ a.	… [$_{\mathrm{VP}}$ ADV t_V NP]	
b.	… NP$_{\mathrm{j}}$ [$_{\mathrm{VP}}$ ADV t_V t_j]	!*

(6)

	input: strong indefinite object	SCOPING
a.	… [$_{\mathrm{VP}}$ ADV t_V NP]	!*
☞ b.	… NP$_{\mathrm{j}}$ [$_{\mathrm{VP}}$ ADV t_V t_j]	

The deeper interest of appealing to OT, however, only comes to the fore when one considers some slightly more complicated variants of (1). Thus, as illustrated in (7) and (8), object-shift is blocked if periphrastic verb forms are used.

(7) a. Auk þess hafa þau alltaf sýnt viðtal við Clinton
í erlendu fréttunum.
Besides have they always shown interviews with C.
in the foreign news.
b. *Auk þess hafa þau viðtal við Clinton alltaf sýnt
í erlendu fréttunum.

(8) a. … [$_{\mathrm{VP}}$ alltaf [$_{\mathrm{VP}}$ sýnt [$_{\mathrm{NP}}$ viðtal við Clinton] …]]
b. *… [$_{\mathrm{AgrOP}}$ [$_{\mathrm{NP}}$ viðtal við Clinton]$_{\mathrm{j}}$ [$_{\mathrm{VP}}$ alltaf [$_{\mathrm{VP}}$ sýnt t$_{\mathrm{j}}$ …]]]

Given what has been said about the interpretation of (1), one might (naively) predict that as soon as periphrastic verb forms are used, the strong indefinite reading can no longer be expressed. However, where the shifted form is missing, the *in-situ* form, here (7a), supports *both* the weak and the strong indefinite reading, that is, it is ambiguous. Clearly, the Scoping constraint can be overruled by some purely syntactic constraint. Vikner (2001, p. 328) takes Licensing, given in (9), to be that constraint.

(9) Licensing

An object must be licensed by being c-commanded by its selecting verb at S-structure.

Licensing captures the fact that object-shift can only take place if the main verb has left VP. This is the case in (1), where the main verb is finite, since finite verbs undergo verb second movement in Icelandic. In (7), on the other hand, the verb second constraint is satisfied by the finite auxiliary, while the participial main verb must stay inside VP. As (8b) shows, the VP-internal participle is unable to c-command the shifted object. Thus, Licensing is violated. In order to yield the desired results, Licensing must outrank Scoping, as shown in (10):

(10) Licensing >> Scoping

Under these assumptions, both readings get assigned the same form, as (11) and (12) demonstrate:

(11)

	input: weak indefinite object	Licensing	Scoping
☞ a.	… ADV V NP		
b.	… NP$_j$ ADV V t$_j$	!*	*

(12)

	input: strong indefinite object	Licensing	Scoping
☞ a.	… ADV V NP		*
b.	… NP$_j$ ADV V t$_j$	!*	

Clearly, due to the use of defeasible constraints, OT provides a very elegant account of this non-trivial interaction between formal and interpretive constraints.[5] On a more intuitive note, it can even be argued that Scoping is responsible for a disambiguating effect partly masked by the workings of Licensing. In that sense, disambiguation would be "the emergence of the unmarked". Taking this line of reasoning more seriously, I would like to propose a more explicit approach to disambiguation. Thus, assume there is a family of disambiguation constraints called "Unambiguous Encoding" (UE). The relevant Icelandic variant of UE would be instantiated as in (13):

(13) Unambiguous Encoding (UE) [Icelandic]:

a. Weak indefinites stay *in situ*.
b. Strong indefinites shift.

Of course, UE could replace Scoping and interact with syntactic constraints like Licensing, yielding the same effects in the domain of Icelandic object-shift. Before I go into the more conceptual discussion of UE, let us have a look at another potential instantiation.

2.2 (In-)Definite theme marking in Tagalog

Consider the following alternation in simple transitive clauses of Tagalog:

(14) a. Bumili ang lalaki ng bigas.
AT-bought T man Th rice
'The man bought (*the) rice.'
b. Binili ng lalaki ang bigas.
ThT-bought A man T rice
'The/A man bought the rice.'

As is well known, verb morphology in Tagalog covaries with the (class of the) thematic role of a designated argument. I will call that argument "trigger", following Schachter's (1993) theory-neutral usage. The trigger is immediately preceded by the "trigger-marker" *ang* (glossed *T*). In (14a) the trigger is an agent, while in (14b) it is a theme/patient. This motivates agent-trigger (*AT*) versus theme-trigger (*ThT*) morphology on the respective verbs. The non-trigger argument is immediately preceded by a default marker *ng*. As can be gathered from the translations, the markers come with a definiteness effect. If we concentrate on theme arguments, we can interpret this as another instance of UE, the formulation of which is given in (15):[6]

(15) Unambiguous Encoding (UE) [Tagalog]:

a. Indefinite theme is *ng*-marked.
b. Definite theme is *ang*-marked.

Adding (16) as the constraint governing trigger morphology and the constraint ranking in (17), we can derive the pattern in (14). This is shown schematically in (18) and (19):

(16) Syn1: *Ang*-markers on NPs correspond to verb morphology.

(17) Syn1 >> UE

(18)

	input: indefinite theme	Syn1	UE
☞ a.	AT-V *ang*-A *ng*-Th		
b.	AT-V *ng*-A *ang*-Th	!*	*
c.	ThT-V *ang*-A *ng*-Th	!*	
d.	ThT-V *ng*-A *ang*-Th		!*

(19)

	input: definite theme	Syn1	UE
a.	AT-V *ang*-A *ng*-Th		!*
b.	AT-V *ng*-A *ang*-Th	!*	
c.	ThT-V *ang*-A *ng*-Th	!*	*
☞ d.	ThT-V *ng*-A *ang*-Th		

Interestingly, as was the case with UE[Icelandic], the disambiguative power of UE[Tagalog] can be masked by formal constraints. Thus, consider Tagalog relative clauses. These obey a "trigger only" condition (cf. Keenan and Comrie, 1977), that is, only triggers can be relativized. Since Tagalog does not employ relative pronouns, the surface effect of this relativization strategy is that no *ang*-marked element appears among the immediate constituents of a relative clause. This is illustrated in (20):[7]

(20)	a.	…lalaking	bumasa	ng	diyaryo
		man-Li	*AT-read*	*Th*	*newspaper*
		'…man who read newspaper'			
	b.	*…lalaking	binasa	ang	diyaryo
		man-Li	*ThT-read*	*T*	*newspaper*
	c.	…diyarong	binasa	ng	lalaki
		newspaper-Li	*ThT-read*	*A*	*man*
		'…newspaper which man read'			
	d.	*…diyarong	bumasa	ang	lalaki
		newspaper-Li	*AT-read*	*T*	*man*

Again one could (naively) predict that definite reference of theme arguments surfacing in relative clauses cannot be expressed. However, examples to the contrary, like the one in (21), have been reported in the literature, for example by Schachter and Otanes (1972, p. 382f.) and Maclachlan and Nakamura (1997, p. 311):

(21) Matalino ang lalaking bumasa ng diyaryo.
intelligent T man-Li AT-read Th newspaper
'The man who read a/the newspaper is intelligent.'

Thus, a purely formal constraint is given priority over an interpretive one. Once more we seem to have a non-trivial case for OT-style constraint interaction. Assume the following to be an appropriate version of the formal constraint responsible for the "trigger only" condition on relatives:

(22) Syn2: Relative-operators are *ang*-marked.

Like Syn1, Syn2 outranks UE. Note also that we can take Syn1 and Syn2 to be "tied". This is shown in (23):

(23) Syn1 <<>> Syn2 >> UE

Such a system straightforwardly derives the fact that definite theme arguments can be *ng*-marked in relative clauses, in violation of UE. The relevant competitions in (25) and (26) are based on agent-relative clauses only, as schematically represented in (24). (The abstract relative operator is represented by "Op".)[8]

(24) [read *A*-Op *Th*-newspaper]

(25)

	input: indefinite theme	Syn1	Syn2	UE
☞ a.	AT-V *ang*-A-Op *ng*-Th			
b.	AT-V *ng*-A-Op *ang*-Th	!*	*	*
c.	ThT-V *ang*-A-Op *ng*-Th	!*		
d.	ThT-V *ng*-A-Op *ang*-Th		!*	*

(26)

	input: definite theme	Syn1	Syn2	UE
☞ a.	AT-V *ang*-A-Op *ng*-Th			*
b.	AT-V *ng*-A-Op *ang*-Th	!*	*	
c.	ThT-V *ang*-A-Op *ng*-Th	!*		*
d.	ThT-V *ng*-A-Op *ang*-Th		!*	

2.3 For unambiguous encoding

The previous two sections have made the empirical case for a family of disambiguation constraints called "Unambiguous Encoding" (UE). Further applications can easily be envisaged.[9] To the extent that this approach is on the right track, it points to a generalization that would be lost if everything were left to individual analyses, like the one of Icelandic object-shift by Vikner (1997, 2001).[10] This is in line with the deeper "functionalist" intuition that disambiguation may be one of the hidden factors underlying (OT) constraints at the syntax/semantics interface. Our approach would make it possible to give a pragmatic foundation for grounding such constraints. The obvious candidate for UE, of course, is Gricean "Avoid Ambiguity" or "Be Perspicuous".[11]

While being attractive in its own right, this approach is further strengthened by current attempts at providing foundations for Gricean maxims within (evolutionary) game theory (cf. Asher, Sher and Williams, 2001; van Rooy, Chapter 8, and forthcoming).

3 The epiphenomenal status of unambiguous encoding

Having stated the case for UE in Section 2, I would like to devote the current section to exploring a number of objections (Section 3.1) and alternatives (Sections 3.2, 3.3) to the UE approach.

3.1 Against unambiguous encoding

First, one may have qualms about the conjunctive nature of UE.[12] Thus a closer look reveals that the violations of UE discussed in Sections 2.1 and 2.2 are due to one clause of UE only, namely, the *b*-clause. As will become clearer later on, this means that the more marked readings can fail to be associated with the more marked forms (strong indefinites may fail to shift; definite themes may fail to be *ang*-marked). However, we have no evidence for the converse, that is, less marked readings do not seem to get associated with more marked forms.[13] This indicates that the *a*-clauses of UE might potentially be eliminated. Alternatively, they could be independent high-ranked constraints, the violability of which would have to be explored.

Second, one may object to UE on the count that it is too explicit. Thus, it is easy to formulate the following inverse form/meaning associations, yielding "Anti-Icelandic" and "Anti-Tagalog" respectively:

(27) UE[Anti-Icelandic]:

a. Weak indefinites shift.
b. Strong indefinites stay *in situ*.

(28) UE[Anti-Tagalog]:

a. Indefinite theme is *ang*-marked.
b. Definite theme is *ng*-marked.

A similar critique has been raised by Zeevat and Jäger (2002) with regard to Aissen's "harmonic alignment" approach, a variant of which will be discussed below in Section 3.3. In fact, this is a general weakness of so-called "input/output" constraints like UE. To the extent that principles of this kind are unwelcome – at least empirically the predicted patterns are unattested – it would be preferable if disambiguation were to fall out "implicitly" as a by-product of the optimization process. A system capable of producing such an effect is Blutner's bidirectional OT, which will be discussed in Section 3.2.

Third, as pointed out by Reinhard Blutner (p.c.), a "pure" version of UE or "Avoid Ambiguity", that is, a version without (language-) specific instantiations, could not operate on the constraint level directly. Instead, it would constrain the input/output relation at a meta-level and thus exceed the power of standard OT.

Finally, it is clear that ambiguity is such an omnipresent feature of natural language that it may be misguided to expect explicit constraints like UE that cut down its power (cf. Zeevat, 2000, p. 245). As will become clearer in the following sections, however, the disambiguation patterns we have been dealing with are of a special kind related to "Horn's division of pragmatic labor", according to which (un)marked meanings get associated with (un)marked forms. It can be argued that a special approach to this kind of disambiguation remains attractive, even if other cases of ambiguity resolution fall into the domain of different (e.g., processing) components.

3.2 Bidirectional OT

The optimization part of OT, as it is commonly conceived, works in the direction of encoding. Thus, the standard procedure is to find the optimal expression for a given semantic input. This can be called "OT syntax" (cf. Anttila and Fong, 2000). Recently, arguments for taking in the decoding perspective have been developed in an enterprise called "OT semantics"

(Hendriks and de Hoop, 2001). Blutner (2000), on the other hand, argues that the proper treatment of interpretation in OT requires both perspectives.[14] The most attractive aspect of Blutner's approach in the current context is that patterns of disambiguation or "partial blocking" directly result from the optimization procedure. This crucially involves an OT variant called "weak bidirectional OT", whose central notion of "super-optimality" can be defined as follows ($<$ = "is more economical/less costly than"):[15]

(29) A form-meaning pair (f,m) is ***super-optimal*** iff $(f,m) \in$ Gen, and

(Q) there is no other super-optimal pair (f',m): $(f',m) < (f,m)$, and
(I) there is no other super-optimal pair (f,m'): $(f,m') < (f,m)$

For a treatment of disambiguation in the sense of UE[Icelandic] and UE[Tagalog], we have to assume the following rankings of forms and meanings:

(30) Icelandic:

a. f_1 = *in situ* (object) [IS] $<$ f_2 = object-shift [OS]
b. m_1 = weak [W] $<$ m_2 = strong [S]

(31) Tagalog:

a. f_1 = *ng* [NG] $<$ f_2 = *ang* [ANG]
b. m_1 = indefinite [I] $<$ m_2 = definite [D]

These rankings can be taken to be induced by markedness constraints. Thus, to account for (30a), it is natural to assume that movement (or dependency formation) deriving object-shift is costly.[16] Likewise, (30b) is justified if we follow van der Does and de Hoop (1998) in claiming that strong readings of indefinites involve type-shifting, an operation which can again be taken to cause additional costs. (31a) follows to the extent that *ang*-marking requires establishing a costly morphological dependency between an argument and the main verb.[17] Finally, we can attribute the ranking in (31b) to the principle that definite reference involves costlier contextual assumptions (e.g., those due to accommodation). On the basis of these rankings and definition (29) we derive two super-optimal form/meaning pairs in each case. These are charted in (32a) and (32b) for Icelandic and Tagalog respectively:

(32) a.

f_1	☞ IS-W	←	IS-S
	↑		↑
f_2	OS-W	←	☞ OS-S
	m_1		m_2

b.

f_1	☞ NG-I	←	NG-D
	↑		↑
f_2	ANG-I	←	☞ ANG-D
	m_1		m_2

Schematically (f_1,m_1) is super-optimal since individually f_1 and m_1 are the most economical forms. Changing f_1 or m_1 could not make the resulting pair more economical. (f_2,m_1) and (f_1,m_2) are not super-optimal because they are "blocked" by the super-optimal pair (f_1,m_1), the former due to (Q) the latter due to (I). Finally, given that (f_2,m_1) and (f_1,m_2) are not super-optimal, (f_2,m_2) is, because (f_2,m_1) and (f_1,m_2) are the only candidates that could have blocked (f_2,m_2) in terms of (I) and (Q) respectively. As a result we automatically derive the patterns in (1)/(5)/(6) and (14)/(18)/(19) without any appeal to explicit constraints like UE. In this sense, super-optimality makes UE epiphenomenal. In fact, Blutner's weak bidirectional OT indicates that Icelandic object-shift and Tagalog (in)definite theme-marking should be looked at as instances of "Horn's division of pragmatic labor", according to which (un)marked forms get associated with (un)marked meanings.

Unfortunately, as it stands, this account appears to be incomplete. Thus, recall that UE can be overruled where periphrastic tenses come into play in Icelandic or relative clause formation in Tagalog. Under the current perspective, this means that when one form, f_2, is lacking, f_1 acquires both readings, m_1 and m_2, that is, f_1 becomes ambiguous. This does not follow from the above assumptions. Instead it is incorrectly predicted that (f_1,m_1) blocks (f_1,m_2), given that (f_1,m_1) < (f_1,m_2). This is shown in (33):

(33) a.

f_1	☞ IS-W	←	IS-S
	m_1		m_2

b.

f_1	☞ NG-I	←	NG-D
	m_1		m_2

In order to remedy this situation, one has to first of all find a way of associating f_1 and m_2. This may actually be done by means of a contextual constraint. Thus, note that in the case of Icelandic object-shift the readings of the indefinite "interviews with Clinton" coincide with different topic/comment structures (cf. Erteschik-Shir, 2001). Simplifying somewhat, we can say that the strong reading is appropriate if the utterance is about interviews with Clinton. The weak one is appropriate when the utterance is about the foreign news. This can be captured in terms of a contextual constraint Top, given in (34):[18]

(34) Top: Weak/Strong indefinites are in/compatible with a "topical" referent.

If Top is more important than the interpretive constraint responsible for the ranking of meanings in (30b) – call it M for concreteness – then m_1 and m_2 can be reranked according to context. Consequently, as shown in (35), f_1 can be associated with either m_1 or m_2:

(35) a.

context: *about(foreign news)*		
f_1	Top	M
m_1		
☞ m_2	!*	*

b.

context: *about(interviews with Clinton)*		
f_1	Top	M
m_1	!*	
☞ m_2		*

Still, in contrast to the earlier OT syntactic approach, Viknerean or UE-based, ambiguity never arises in this system. What's more, while appeal to Top can render the pair (f_1,m_2) super-optimal in the right context, there is no way to prevent (f_2,m_1) from being super-optimal in that context as well. While in OT-syntax f_2 is simply filtered out during evaluation, weak bidirectional OT keeps f_2 as a candidate form, be it the dispreferred one. By the logic of super-optimality, then, f_2 associates with the dispreferred reading and thus, in the context of (35b), (f_2,m_1) survives evaluation as well.[19] It therefore looks as if further improvement could only come from stronger assumptions about the OT architecture, something that would affect the elegance of the original approach.[20]

3.3 Harmonic alignment and local conjunction

Let me finally turn to a second way of reducing UE. This requires that we look at Icelandic object-shift, which for the sake of brevity I concentrate on here, as an instance of "differential object-marking" in the sense of Aissen (2000). Aissen (2000) shows how to derive an ordered layer of intricate case-marking systems in terms of the alignment of two prominence scales, the Relational Scale, (36), and the Definiteness Scale, (37) (> = "is more prominent than"):

(36) Relational Scale: Su(bject) > Ob(ject)

(37) Definiteness Scale:
Pronoun(Pro) > Name(PN) > Definite(Def) > IndefiniteSpecific(Spec) > NonSpecific(NSpec)

Restricted to objects, harmonic alignment (cf. Prince and Smolensky, 1993) yields a "harmony scale" (38a), which is reinterpreted as a constraint hierarchy, given in (38b) (⊃= "is more harmonic than"):

(38) Harmonic Alignment:

a. Ob/NSpec ⊃ Ob/Spec ⊃ Ob/Def ⊃ Ob/PN ⊃ Ob/Pro
b. *Ob/Pro ≫ *Ob/PN ≫*Ob/Def ≫ *Ob/Spec ≫ *Ob/NSpec

Thus, by (38a), non-specific objects are more harmonic than specific ones, specific ones more harmonic than definite ones and so on. Conversely (38b) says that being a pronominal object is penalized more severely than being an object proper noun, which itself is penalized more severely than being a definite object and so on. For this hierarchy to have a bearing on object-shift, it has to be "locally conjoined" with a constraint enforcing object-shift. Let's assume *IS to be adequate for that purpose:[21]

(39) *IS: Avoid *in-situ* positions.

Local conjunction yields the constraint hierarchy in (40):

(40) *Ob/Pro & *IS ≫ (Stay ≫) [Swedish]
*Ob/PN & *IS ≫
*Ob/Def & *IS ≫
*Ob/Spec & *IS ≫ (Stay ≫) [Icelandic]
*Ob/NSpec & *IS

According to (40), pronominal objects *in situ* will be penalized more severely than object proper nouns *in situ* and so on, or, the other way round, shifting pronominal objects is more important than shifting object proper nouns and so on. As (40) also indicates, different languages can be defined by the place at which Stay, that is, a constraint banning the application of movement, is inserted. We can thus derive the fairly restrictive Swedish (or Mainland Scandinavian) pattern, where only pronouns may undergo object-shift, by inserting Stay high. The Icelandic pattern results from demoting Stay to a fairly low position, assuming in addition that the break-off point between specific and non-specific reference coincides with the break-off point between "strong" and "weak" readings.[22] Thus, as before, specific/strong indefinite objects will undergo object-shift while non-specific/weak ones stay *in situ*. To complete the picture for periphrastic tenses, we only have to rank Licensing on top, as is shown in (41):[23]

(41) Licensing ≫ *Ob/Pro & *IS ≫ … ≫ Stay ≫ *Ob/NSpec & *IS

That UE can be dissolved this way is highly significant. Clearly, the major ingredients of Aissen's approach, namely, the Relational Scale, the Definiteness Scale, and their (harmonic) alignment are candidates for natural language universals.[24] The conjunctive nature of UE is likewise undone, insofar as there is an independent constraint about strong ("specific") indefinites that outranks the one about weak ("nonspecific") indefinites. This is necessary for Stay to be able to intervene and bring about the differential behavior of weak and strong indefinites.

However, as has already been pointed out by Zeevat and Jäger (2002) (cf. Aissen, 1999, p. 703), an Aissen-style approach retains the problematic capability of defining things like "Anti-Icelandic". This is due to "local conjunction". Thus, if we exchange Stay (= *Move) and *IS in (40), we arrive at the picture in (42):

(42) *Ob/Pro & Stay ≫ (*IS ≫) [Anti-Swedish]
*Ob/PN & Stay ≫
*Ob/Def & Stay ≫
*Ob/Spec & Stay ≫ (*IS ≫) [Anti-Icelandic]
*Ob/NSpec & Stay

(42) defines languages where only pronominal objects stay *in situ* while all other objects shift ("Anti-Swedish") and languages where all kinds of objects stay *in situ* except for non-specific/weak ones ("Anti-Icelandic"). The same "Anti-Horn" pattern would actually result if we were to alter the direction of alignment. Without a purely markedness-based optimization procedure like the one in Blutner's bidirectional OT, there is no deeper reason for why (non-)prominent elements on the Relational Scale prefer association with (non-)prominent elements on the Definiteness Scale. Inverse alignment would yield the "harmonic scale" and constraint hierarchy in (43), from which "Anti-Icelandic" (and "Anti-Swedish") can again be derived

straightforwardly:

(43) a. Ob/Pro ⊃ Ob/PN ⊃ Ob/Def ⊃ Ob/Spec ⊃ Ob/NSpec
b. *OB/NSPEC ≫ *OB/SPEC ≫ *OB/DEF ≫ *OB/PN ≫*OB/PRO

A more subtle point concerns the heterogeneous nature of the Definiteness Scale, which seems to include expression-type categories at the upper end as well as semantic categories at the lower end. Of course, disambiguation, which has been my concern in this chapter, only makes sense with regard to semantic categories that get encoded by formal categories.[25] It therefore seems that disambiguation is (at best) a subcase of "differential marking" and may still deserve special attention of the kind it receives in weak bidirectional OT or the UE approach.

Finally, like Vikner's SCOPING, my UE, and Zeevat's PARSE MARKED (Zeevat, 2000), the Aissen-style system must have recourse to input/output constraints, namely, (the local conjunction of) *OB/SPEC&*IS as well as *OB/NSPEC&*IS, in order to properly associate (un)marked forms and (un)marked meanings. This is one of the features that only Blutner's weak bidirectional OT avoids.

4 Conclusion

We have seen that Icelandic marks the difference between weak and strong readings of indefinites by the (non-)application of object-shift. Tagalog encodes the difference between indefinite and definite theme arguments by means of *ng*- versus *ang*-marking. I have argued that these facts are indicative of a deeper "functionalist" drive of grammars toward disambiguation. More specifically, these patterns can be seen as instances of Horn's "division of pragmatic labor" according to which (un)marked forms associate with (un)marked meanings. We have also seen that these "Horn patterns" can be masked by purely formal constraints that make the marked forms unavailable. Thus, object-shift in Icelandic is blocked when perphrastic tenses are used. Tagalog *ang*-marking cannot take place inside relative clauses. Under these circumstances, the unmarked forms ambiguously encode both meanings. Taken together, these facts can be interpreted as establishing patterns of "partial disambiguation" or "partial Horn patterns", where pure disambiguation or pure Horn patterns constitute "the emergence of the unmarked". I have argued that OT is particularly well suited to account for this, given most crucially that it permits defeasible constraints.

In order to deepen our understanding of this success, I have compared and contrasted four different OT approaches to the above facts, three OT syntactic ones and one bidirectional approach. Clearly, only the OT syntactic systems are successful in partially eliminating marked forms and thus deriving partial Horn patterns. Thus, even if the weak bidirectional framework of

Blutner (2000), enriched with context constraints, is able to conditionally associate unmarked forms with marked meanings, it cannot prevent marked forms being assigned the unmarked meanings under exactly these conditions.[26]

Conversely, only weak bidirectional OT fully satisfactorily accounts for emerging pure Horn patterns, requiring just independent complexity measures for forms and meanings. Given these, (un)marked forms are associated with (un)marked meanings due to an optimization procedure that evaluates the cost of form/meaning pairs. The OT syntactic approaches, on the other hand, have to rely on (defeasible) input/output constraints in order to deal with pure Horn patterns. Their weakness consists in the fact that they can easily be rephrased such that "Anti-Horn patterns" (unmarked forms associate with marked meanings, marked forms with unmarked meanings) result instead.[27]

Another line of inquiry pursued in this chapter concerns the relation of OT(-constraints) to pragmatics. Thus, Blutner's weak bidirectional OT is a direct recast of Gricean maxims in their reconstructed versions provided by Horn and Levinson (cf. Blutner, 2000). This opens up the possibility of grounding OT in pragmatics. Moreover, van Rooy (Chapter 8, and forthcoming) has shown how to provide game-theoretical foundations for weak bidirectional OT. More specifically, if assumptions from evolutionary game-theory are combined with Lewisian "signalling games", Horn patterns turn out to correspond to dominant evolutionarily stable states. Among OT syntactic approaches, such a foundation may be most easily replicable, if, as I have argued in Section 2, OT is enriched with a family of disambiguation constraints called "Unambiguous Encoding" (UE), which directly captures Horn patterns in their language-specific instantiations. UE may then be taken to be grounded in Gricean "Be Perspicuous" or "Avoid Ambiguity". On a speculative note, I would like to add that "Be Perspicuous" itself may be evolutionarily grounded on the basis of Axelrod's (1984, p. 54) result that "clarity" (≈ intelligibility) is one of the prerequisites for successful cooperation.

Compared to the two remaining linguistically more respectable OT syntactic approaches, the UE system suffers from its "shallowness" as far as the encoding of grammatical principles is concerned. Thus, Vikner's (1997, 2001) account of Icelandic object-shift is built on the more widely applicable input/output constraint Scoping. However, an independent account of Tagalog would have to be provided, leaving little hope for capturing the "functionalist" flavor of conditional Horn patterns in a unified way. This point may weigh in favor of an Aissen-style approach in terms of "harmonic alignment" and "local conjunction" (Aissen, 1999, 2000), which could render UE epiphenomenal. What is attractive about such a system is its appeal to a universal constraint hierarchy derived from the Relational Scale and Definiteness Scale, themselves potential natural language universals. Over and above allowing for a much more unified approach to Horn patterns as they emerge in our Icelandic and Tagalog cases, an Aissen-style approach may also be amenable to reconstruction in frequentistic terms, as has been outlined by Zeevat and Jäger (2002) and further worked out on the basis of a bidirectional learning algorithm by Jäger (Chapter 11).

Acknowledgements. For comments and suggestions I would like to thank Peter Ackema, Reinhard Blutner, Cleo Condoravdi, Helen de Hoop, Gerhard Jäger, Paul Kiparsky, Jonas Kuhn, Jason Mattausch, Torgrim Solstad, Ralf Vogel, Henk Zeevat, and two anonymous reviewers, as well as audiences at the GGS meeting in Bochum, the Stanford Linguistics Colloquium, the third Szklarska Poreba meeting on the "Roots of Pragmasemantics", the workshop on "Optimality Theory and Pragmatics" in Berlin, the Pionier Workshop on "Variation in Form versus Variation in Meaning" in Nijmegen, and the ESSLLI workshop on "Pragmatics in OT" in Trento. Special thanks go to Florian Schäfer for input to and discussion of the analysis of Tagalog. Common disclaimers apply.

Notes

1. For recent debates on functionalism and OT, see (comments on) Haspelmath's contribution to *Zeitschrift für Sprachwissenschaft* 18.2 (1999), and the exchange between Newmeyer, Aissen and Bresnan in *Natural Language and Linguistic Theory* 20 (2002).
2. From the perspective of phonology, the issue of grounding has been explicitly addressed by Kager (1999, p. 11): "Phonological markedness constraints should be phonetically grounded in some property of articulation or perception."
3. I abstract away from another more restricted reading of (1b), paraphrasable as *If they are shown at all, interviews with Clinton are shown in the foreign news*. Thanks to Jason Mattausch and Torgrim Solstad for making me aware of this. Also, as noted by Vikner (2001), the limitation of (1a) to reading (3a) may not be fully strict but rather a matter of strong preference. If so, the issues addressed in this chapter will have to be eventually recast in frequentistic terms such as outlined in Zeevat and Jäger (2002) and Jäger (Chapter 11).
4. I have slightly adapted Vikner's formulation of this constraint.
5. As already noted by Vikner (2001), alternative non-OT accounts of these Icelandic facts make use of defeasible constraints as well. This is perhaps most conspicuous for the earliest minimalist account by Bobaljik, going back to 1994 (see Bobaljik, 2002). There an input/output constraint called "Minimize Mismatch" assimilates LF and PF representations in a way similar to SCOPING, as long as it isn't overruled by purely syntactic principles.
6. This effect can be suspended if complex NPs bearing cardinal determiners are used (cf. Adams and Manaster-Ramer, 1988). Also, the effect may be a strong preference rather than fully strict (cf. Himmelmann, forthcoming). See also Note 3.
7. *Li* glosses the "linking morpheme" *ng* encliticizing between a modified head and its modifier.
8. The same effect shows up in the trigger-less "recent past" construction of Tagalog. There being no possibility of *ang*-marking in these structures, *ng*-marked themes can ambiguously be both indefinite and definite. Again, UE is overridden.

9. The following phenomena appear straightforwardly amenable to a UE approach: Elative/partitive choice in Finnish (Anttila and Fong, 2000), focus marking in African languages (Hyman, 1984), specificity and direct object marking in Hindi (Mohanan, 1994), and instrumental object marking in Greenlandic Eskimo (Fortescue, 1984).
10. A UE-less OT analysis of the Tagalog facts can be built fairly directly on the basis of Maclachlan and Nakamura (1997), whose account crucially involves ranked defeasible constraints.
11. This line of thought picks up earlier ideas by people like Hyman (1984, p. 73) on "the harnessing of pragmatics by a grammar", or Levinson (1987a, p. 420) on "frozen pragmatics".
12. If we allow variable α to range over positive vs. negative specification [+/−], the following may be more elegant formulations of UE:
    (i) UE[Icelandic]: [α strong] → [α shifted]
    (ii) UE[Tagalog]: [α definite_theme] → [α *ang*-marked].
13. As one anonymous reviewer points out, such evidence may be found in the lexicon, where for example "certainly", "sure", or "fine" can neutrally replace "yes".
14. See Zeevat (2000) for some critical evaluation of these alternatives.
15. Cf. Blutner (2000, p. 204fn.7). As shown by Jäger (2002), this kind of recursive definition yields proper results if the ordering relation is transitive and well-founded. Beaver and Lee (Chapter 6) provide a systematic overview of how various OT systems deal with form/meaning (mis)matches. Further discussion and a game-theoretical reconstruction of "super-optimality" can be found in Dekker and van Rooy (2000). Naming the clauses Q and I, of course, goes back to Horn's and Levinson's reconstruction of Gricean maxims.
16. As pointed out by Bresnan (2001b), this builds on the assumption that "unmarked" structures coincide with D-structures in generative syntax. If, as suggested by one anonymous reviewer, object-shift were analyzed as displacement of a larger constituent carrying along (i.e. "pied-piping") VP and the object it contains, this ranking of forms would have to be recast accordingly. Thus, the shifted form could be considered as "marked", given that *more* material is pied-piped across the medial adverb.
17. This assumption may be surprising from the perspective of theories equating *ang*-marking with subjecthood (cf. Kroeger, 1993). However, such an approach is not uncontroversial (cf. Maclachlan and Nakamura, 1997; Schachter, 1993). A more comprehensive OT treatment of Tagalog must await further research.
18. As pointed out by one anonymous reviewer, constraint (34) is just a recast of SCOPING if the strong/weak distinction can be exhaustively captured in such information-structural terms (and a proper discourse-semantic framework is adopted). Here, the strategic point of (34) is to give m_2 an advantage over m_1 in context. Providing an adequate theory of context-based interpretation in OT is beyond the scope of this chapter.
19. See Beaver and Lee (Chapter 6) for a more general illustration of this effect.
20. Zeevat (2000) sketches a modified version of bidirectional OT that would work here. This approach asymmetrically orders OT syntax before OT semantics, so that certain form/meaning pairs can be filtered out by OT syntax alone. In the case of Icelandic object-shift, a high-ranked constraint like LICENSING could eliminate f_2 where periphrastic verb forms are used. Only (f_1,m_1) and (f_1,m_2) would enter OT semantics, and would thus be potentially well-formed outputs.

However, in order to derive the original pattern of "partial blocking", that is, the coexistence of (f_1,m_1) and (f_2,m_2), an input/output correspondence constraint similar to SCOPING or UE has to be assumed. Zeevat (2000, p. 258) suggests the more general constraint PARSE MARKED, which associates marked meanings with marked forms, and thus explicitly states what is an automatic result of Blutner's weak BiOT.

Beaver (to appear) sketches a two-stage bidirectional architecture where a constraint, *BLOCK, sensitive to markedness of candidates as computed in the first stage, brings about "partial blocking" in the second round of optimization. Since this constraint can itself be outranked, for example, by LICENSING, the masking effect can likewise be derived. See also Jäger (Chapter 11).

21. To the extent that Aissen's Case-marking constraint *Ø can be motivated by "Minimization of Perceptual Confusion" (Aissen, 2000), *IS may have to be given a similar function, which it would seem to have within radical approaches like Beghelli and Stowell (1997) and Hornstein (2001).
22. The full pattern of languages that (40) is suggestive of may be found if scrambling languages are taken into consideration as well.
23. The approach developed in Aissen (1999, 2000) is clearly relevant for Tagalog as well. However, given the controversial status of grammatical relations in that language, it may be advantageous to analyze it in terms of "differential theme marking". This would require alignment of the Definiteness Scale with the Thematic Scale (Agent > Theme).
24. It is one of the crucial ideas behind harmonic alignment that the resulting constraint hierarchies cannot be reranked, unlike standard OT constraints, but express a universal ranking.
25. According to Aissen (2000), the semantic intuition behind the Definiteness Scale is "the extent to which the value assigned to the discourse referent introduced by the noun phrase is fixed". Whether this straightforwardly corresponds to the range of categories like "pronoun" and "proper noun" is doubtful.
26. Strictly speaking, this is not what happens. Rather, markedness of meanings is taken to vary with context.
27. This is not to say that "Anti-Horn patterns" are beyond a bidirectional approach. In fact, the following definition of "super-optimality*" would generate them.

(i) A form-meaning pair (f,m) is ***super-optimal**** iff $(f,m) \in$ Gen, and, if there are pairs $(f',m) \in$ Gen or $(f,m') \in$ Gen $(f' \neq f;\ m' \neq m)$, then

(Q*) there is a non-super-optimal* pair (f',m): $(f',m) < (f,m)$, or

(I*) there is a non-super-optimal* pair (f,m'): $(f,m') < (f,m)$

8
Relevance and Bidirectional Optimality Theory

Robert van Rooy

1 Introduction

According to optimality theoretic semantics (e.g., Hendriks and de Hoop, 2001), there exists a gap between the semantic representations of sentences and the thoughts actually communicated by utterances. How should this gap be filled? The obvious answer (Grice, 1957) seems to be that the hearer should recognize what the speaker thinks that the listener understands. Because this depends in turn, in a circular way, on what the listener thinks that the speaker has in mind, a game-theoretical framework seems natural to account for such situations. Intuitively, what goes on here is a game between a speaker and a hearer, where the former chooses a form to express the intended meaning, and the latter chooses a meaning corresponding to the form. Blutner's Bidirectional Optimality Theory (OT), based on the assumption that both speaker and hearer optimize their conversational actions seems perfectly suitable to implement this. But how can a hearer recognize the speaker's intentions? Gricean pragmatics (1975) suggests that she can do so by assuming that the speaker is cooperative and thus obeys the conversational maxims. Sperber and Wilson (1986) have suggested that these four conversational maxims can be reduced to the single principle of optimal relevance. In this chapter I will discuss how far this can be done. I will argue that conversation involves resolving one of the participants' decision problems. After discussing bidirectional optimality theory I will show how decision theory can be used to determine the utility of an interpretation in a mathematically precise way. Then I will discuss how this formal notion of utility, in combination with bidirectional optimality theory, can account for a number of conversational implicatures and how it relates to: (i) Sperber and Wilson's psychologically inspired notion of cognitive relevance; (ii) the Stalnakerian assertability conditions; (iii) the Gricean maxims of conversation; and (iv) the so-called Q- and I-principles of neo-Gricean pragmatics (Horn, 1984; Levinson, 2000).

2 Bidirectional optimality theory

Optimality theory (OT) assumes that a linguistic form should be interpreted in the *optimal* way. The crucial insight behind Blutner's (2000) bidirectional OT is that for the *hearer* to determine what the optimal interpretation of a given form is, he must also consider the *alternative expressions* the *speaker* could have used to express this meaning/interpretation. One way to implement this idea is to say that we not only require that *the hearer* finds the optimal meaning for a given form, but also that *the speaker* expresses the meaning he wants to communicate by using the optimal form. Thus, what is optimal is not just meanings with respect to forms, but rather form-meaning pairs. According to bidirectional OT the form–meaning pair $\langle f,m\rangle$ is **optimal** iff it satisfies both the *S* principle (i.e., is optimal for the *speaker*) and the *H* principle (i.e., is optimal for the hearer):[1,2]

(*S*) $\neg\exists f': \langle f',m\rangle \in H \;\&\; \langle f,m\rangle < \langle f',m\rangle$

(*H*) $\neg\exists m': \langle f,m'\rangle \in S \;\&\; \langle f,m\rangle < \langle f,m'\rangle$

To turn the above definition of optimality into a predictive formalism, we have to know several things: (i) what are the alternative forms? (ii) what are the alternative meanings? and (iii) how should we interpret the ordering relation $<$?

In Blutner (1998, 2000) no restrictions are laid down on what alternative expressions/forms to take into account. Blutner (1998) proposes to let the alternative meanings be Carnapian *state descriptions*. The ordering relation is defined in terms of a *cost*-function, defined in turn on the *complexity* of the forms and the *conditional informativity* of the meanings. The cost of a form-meaning pair $\langle f,m\rangle$, $c(\langle f,m\rangle)$, is then $compl(f) \times \inf(m/[[f]])$, where $compl(f)$ measures the complexity of form f; $[[f]]$ is the 'semantic' meaning of f; and $\inf(m/[[f]])$ measures the surprise that m holds when f is true.[3] I will sometimes call $\inf(m/[[f]])$ the *surprisal* that m holds if $[[f]]$ is true. If f is a sentence like "John said hello to a secretary", we could assume that this gives rise to two interpretations: m, where the secretary is *female*, and m', where the secretary is *male*. Because secretaries are normally female, it holds that $P(m/[[f]]) > P(m'/[[f]])$, that is, m is a more likely given $[[f]]$ than m' is, and thus $\inf(m/[[f]])$ is lower than $\inf(m'/[[f]])$, $\inf(m/[[f]]) < \inf(m'/[[f]])$. The ordering relation between form–meaning pairs is then defined as expected: $\langle f,m\rangle$ is preferred to $\langle f',m'\rangle$ iff the cost of the former is smaller than the cost of the latter, that is, $\langle f,m\rangle > \langle f',m'\rangle$ iff $c(\langle f,m\rangle) < c(\langle f',m'\rangle)$. Thus, in particular, $\langle f,m\rangle > \langle f,m'\rangle$ iff m is a more likely, or stereotypical, interpretation of f than m'' is. Blutner notes that by using this implementation he comes close to implementing Zipf's (1949) idea that interpretation can be seen as a balance of, on the one hand, the force to minimize the *speaker's effort* by preferring forms with a lower complexity, and, on the other, the force to minimize the *hearer's effort* by selecting the worlds that minimize the (conditional) surprise given the semantic meaning of the expression. What is the enriched, or preferred, meaning of sentence *f*? It is the union of meanings *m* such that $\langle f,m\rangle$ satisfies both the *S* and *H* principles. By using this mechanism, Blutner (1998) claims to be able to account for scalar and clausal implicatures classically accounted for in terms of Grice's maxim of Quantity, also known as *Q*-inferences, for the fact that sentences typically get interpreted in *stereotypical* ways (known as *I*-inferences in neo-Gricean pragmatics), and for Horn's (1984) division of pragmatic labor.

2.1 *Q*-inferences

Blutner's bidirectional OT accounts for classical quantity implicatures if we assume that the alternative meanings are worlds. Let's look at the *scalar* implicature derivable from $B \vee C$ that $B \wedge C$ is false and the *clausal* one that *B* and *C* are both possible. Let us assume we have four relevant worlds: w_0 where neither *B* nor *C* are true; w_1 where only *B* is true; w_2 where only *C* is true; and w_3 where both are true. Because $\inf(w/[[B \vee C]])$ has the same value for each world *w* in which '$B \wedge C$' is true (or so let us assume), '$A \vee B$' could be interpreted as $\{w_1,w_2,w_3\}$, as far as the *H*-principle is concerned. However, w_3 is not optimal for the speaker because there is an alternative expression, '$B \wedge C$', such that the surprisal that w_3 holds after learning that this alternative expression is true is smaller than the surprisal that w_3 holds after learning that $B \vee C$ is true: $\inf(w_3/[[B \wedge C]]) < \inf(w_3/[[B \vee C]])$. As a result, '$\mathrm{B} \vee \mathrm{C}$' gets the exclusive interpretation: $\{w_1,w_2\}$. Notice that Blutner's bidirectional OT accounts both for the intuition that from the assertion '$B \vee C$' we conclude that '$B \wedge C$' is not true, that is, the *scalar* implicature, and for the *clausal* implicature that $\Diamond B$, $\Diamond\neg B$, $\Diamond C$, and $\Diamond\neg \mathrm{C}$ are all true. Notice that although the *S* principle *blocks* world w_3 from being 'part' of the meaning of $B \vee C$, this blocking is due to the conditional surprise that orders interpretations, and is *not* due to the fact that there is an alternative *cheaper form* that could express this interpretation/meaning. Blocking, in this case, is thus due to the ordering of meanings, which can depend on the expression being used. This analysis of blocking will be important in Section 6 of this chapter. In the next subsection, however, we will see that bidirectional OT also accounts for blocking due to the existence of more costly alternative expressions.

2.2 *I* inference and Horn's division of labor

Now we will see how, due to the *H*-principle, sentences will be interpreted in stereotypical ways, and, due to the interaction of the *S*- and *H*- principles, marked expressions typically get a marked interpretation. Taken together this pattern is known as Horn's division of pragmatic labor. To illustrate, consider the following well-known example:

(1) a. John stopped the car.
    b. John made the car stop.

Let us assume that both sentences are semantically true if John stopped the car either in a stereotypical way, m_{st}, or in an unusual way, m_u. In that case we typically interpret (1a) as meaning stereotypical stopping, while (1b) as non-stereotypical stopping. Blutner (1998) shows that this is predicted correctly from the interaction of the *S*- and *H*-principles: In case we learn that either (1a) or (1b) is true, the informativity, or surprisal, of m_{st} is smaller than the informativity of m_u, because the former's probability is higher. Because the complexity of (1b) is not smaller than the complexity of (1a), the sentence (1a) is interpreted as m_{st}. Thus, Blutner (1998) accounts for the intuition that sentences typically get the most plausible, or stereotypical, interpretation. To show that the marked form (1b) gets a marked meaning, notice that the interpretation m_{st} is blocked because there is an alternative expression that could express m_{st} in a less complex way. Due to the interaction of the *S*- and *H*-principles, the unmarked (1a) will get the stereotypical interpretation, while the marked (1b) will get the non-stereotypical interpretation.

2.3 OT and constraints

Although bidirectional OT has become rather popular recently to account for certain linguistic data (e.g., Blutner, 2000; Zeevat, 1999, 2000; Aloni, 2001; Krifka, 2002), the specific way in which Blutner (1998) implemented the theory as I presented above has not been taken up. It is not assumed anymore that the form–meaning pairs are ordered in terms of an abstract cost-function. In particular, the idea is given up that the possible meanings of utterance *B* are ordered by the function $\inf(\cdot/[[B]])$ so as to minimize the hearer's effort to interpret. Instead, the analyses are based on Jäger's (2002) proposal to relate bidirectional OT more closely with standard OT approaches: derive the ordering relation between form–meaning pairs from a system of more specific ranked OT constraints, some of which are relevant only for ordering forms, others only for ordering meanings. A number of constraints for ordering meanings are very specific, other are more general and closely related with the assertability constraints of Stalnaker and conversational maxims of Grice.

This new way of doing bidirectional OT opens up many possibilities. But there is also a danger: if one can invent any OT constraint as long as it helps to describe the facts, it is not clear to what extend OT is still explanatory. Remember that Blutner's formulation of bidirectional OT was motivated by the reduction of pragmatics to Zipf's general principle of minimizing speaker's and hearer's effort.[4] The main goal of this chapter is to show how a number of specific OT constraints used in the literature to account for semantic/pragmatic phenomena can be motivated by, or reduced to, very general principles.

3 Bidirectional OT: prospects and problems

We saw that in Blutner's original statement of bidirectional OT the meanings are ordered in terms of one very simple general function: conditional informativity. In this section I want to show both the strength and limits of using this function. In Section 3.2 I will argue that Blutner's use of the informativity function gives rise to a number of problems. These problems will motivate us to look for an alternative general function for ordering meanings. Before we come to that, however, I will argue for the strength of Blutner's informativity function: showing that a number of OT constraints proposed to account for some specific phenomenon can be reduced to this one function. The phenomenon to be discussed is *anaphora resolution* and the theory that was made to account for it is *centering theory*.

3.1 Centering in bidirectional OT

Centering theory is a theory designed to make predictions about anaphoric resolution and the interpretational coherence in discourses. The theory was originally stated by Grosz, Joshi and Weinstein (1983) in a procedural way and has recently been given an attractive optimality theoretical *declarative* reformulation by Beaver (to appear).[5] The original procedural implementation makes use of two rules (called Rule 1 and Rule 2) which Beaver reduces to three violable OT constraints ordered in a hierarchical way.[6] I will show in this section how both the Beaverian constraints *and* the ordering between them follow from Blutner's (1998) original statement of bidirectional OT. This derivation crucially relies on a very similar derivation of the rules of original centering theory proposed by Hasida, Nagao and Miyata (1995). Although my derivation will just be a recoding of theirs in optimality theoretical terms, the derivation is still worth going through, because it shows how specific constraints used in OT can be motivated independently by an ordering relation between form–meaning pairs that is based on a very abstract and general economically based function that orders meanings.

3.1.1 Centering theory

The crucial notions of centering theory are the following:

- C_F^n = *forward looking centers*, the semantic entities referred to in the *n*th sentence in the discourse. They are ranked according to their salience, specified as *grammatical obliqueness*. Ranking is determined by the grammatical functions of the referring expressions in the utterance: (subject > direct object > indirect object > other complements > adjuncts).

- C^n_p = *preferred center of* n = highest ranked element of C^n_F.
- C^n_B= *backward looking center*: the highest ranked element in C^{n-1}_F that is referred to in the *n*th sentence.

Centering theory is now based on two very simple ideas: First, that if a pronoun is used in an utterance, its preferred referent is the backward-looking center of this utterance, called the *topic* of the previous utterance by Beaver (to appear). Beaver notes that this idea (known as *Rule 1*) doesn't have to be stated conditionally once we adopt the OT framework, for now constraints are allowed to be violated. Beaver's constraint, PRO-TOP, to capture Rule 1 of centering theory simply says: The topic must be pronominalized. The second idea of centering theory is that it is assumed that a discourse is more coherent when the topic remains constant, that is, when for each utterance its backward-looking center is the same as that of its previous utterance. This means that a discourse is maximally coherent (as far as anaphoric reference is concerned) if for each utterance n it holds that $C^n_B = C^{n-1}_B$ and $C^n_B \neq C^n_P$.

To illustrate, consider the following discourse:

(2) a. He_1 saw Jack_2 in the park_3.
    b. He_4 stopped his car_5.

The three discourse entities/referents referred to in (2a) are DR_1 (He); DR_2 (Jack) and DR_3 (the park). DR_1 is the center (the C_B of (2a) and also the preferred next center (the C_F) of (2a) and thus the backward center of (2b). Semantically speaking, the pronoun *he* in (2b) could refer back to both DR_1 and DR_2. Giving the centering theoretical preference, however, it is predicted that it will refer to DR_1.

In the above example none of the centering constraints was violated. But what if one or more of these conditions is not satisfied? Which violation is less dramatic than others? According to *Rule 2* of centering theory, transitions are preferred in the following ordering: CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT, where these names have the following denotations:

CONTINUE: $C^n_B = C^{n-1}_B$ and $C^n_B = C^n_P$
RETAIN: $C^n_B = C^{n-1}_B$ and $C^n_B \neq C^n_P$
SMOOTH-SHIFT: $C^n_B \neq C^{n-1}_B$ and $C^n_B = C^n_P$
ROUGH-SHIFT: $C^n_B \neq C^{n-1}_B$ and $C^n_B \neq C^n_P$

Beaver (to appear) notes that the ordering on these transitions can be captured straightforwardly when we assume that $C^n_B = C^{n-1}_B$ and $C^n_B = C^n_P$ are both separate optimality theoretical constraints, dubbed COHERE and ALIGN, respectively, and assume that COHERE is more important than ALIGN.

By assuming in addition that the constraint PRO-TOP is more important than COHERE (and thus also than ALIGN), Beaver's OT reformulation (called COT) also captures the centering theoretic claim that their Rule 1 is more important than their Rule 2.

3.1.2 Deriving the constraints

Although centering theory is normally seen as being purely *descriptive* in that it tries to predict pronoun resolution adequately, at least according to Beaver (to appear) its original motivation was economic in nature:

> One of the driving forces behind early Centering proposals of Joshi and associates was the idea that speakers choose forms which minimize processing costs to hearers. COT models the fact that it may be cheaper in the long-run to use a form which is in the short-term relatively expensive. For instance, a speaker may choose a form in which the topic is not in subject position because it will reduce the costs incurred by a *following* sentence in which a topic shift is needed.
>
> (Beaver, to appear, p. 83)

By assuming that speaker's and hearer's try to minimize their effort, Blutner's bidirectional OT can be seen as a theory of rational language use. This suggests that we should be able to justify the rules of centering theory, or the Beaverian constraints and orderings between them, in terms of Blutner's general formalization of his theory. Following Hasida, Nagao and Miyata, I will suggest that this indeed can be done.

PRO-TOP. The constraint PRO-TOP only has an effect in Beaver's COT if the sentence contains pronouns. In that case it demands that one of them must refer to the backward-looking center. To derive this constraint, let us assume that there is no lighter (anaphoric) expression than a pronoun. It follows from bidirectional OT that this pronoun must thus refer to the best possible meaning. Assume now that the semantic meaning of a pronoun is underspecified, and can be interpreted as any of the elements of the set of forward-looking centers of its previous utterance. Let the forward-looking centers of utterance $n-1$, C_F^{n-1}, be the list $[a, b, c]$ with $C_B^n = a$. In that case we can assume that the semantic meaning of a pronoun in utterance n should be $\{a, b, c\}$. The elements of this list, however, are ordered by salience. In particular, the most probable referent of a pronoun in the nth utterance is its backward-looking center, that is, a. Thus, $\inf(a/\{a, b, c\})$ is smaller than both $\inf(b/\{a, b, c\})$ and $\inf(c/\{a, b, c\})$. Thus, the backward-looking center, that is, a, will be the best meaning, and, by bidirectional OT, will thus be the interpretation of the lightest anaphoric expression (a pronoun). So we see that PRO-TOP follows straightforwardly from Blutner's bidirectional OT.

COHERE. The constraint COHERE is satisfied iff $C_B^n = C_B^{n-1}$.[7] Notice that this constraint can only be violated when the highest-ranked element of C_F^{n-1} is not the same as C_B^{n-1}. In combination with PRO-TOP, this means that the highest-ranked element of C_F^{n-1} could not be referred to in the $n-1$th utterance by a pronoun. Because the use of a pronoun is shorter, and requires less effort, than the use of a proper name or full description to refer to an object, Blutner's bidirectional OT predicts that it is better (for the $n-1$th utterance) to not violate the COHERE constraint.

ALIGN. The constraint ALIGN is satisfied iff $C_B^n = C_P^n$. The reason why bidirectional OT prefers this constraint not to be violated is very similar to the reason why it prefers COHERE to be satisfied, but now related to the *n*th utterance.

From the above derivations of the Beaverian constraints out of bidirectional OT, we can also deduce that PRO-TOP is more important than the other constraints. To derive COHERE for instance, we referred to PRO-TOP, but not the other way around. Moreover, Beaver's ranking between COHERE and ALIGN can be understood also: a violation of COHERE is worse than a violation of ALIGN because the former violation leads to more effort in the $n-1$th utterance, while a violation of ALIGN can only have an effect in the *n*th utterance. In fact, a violation of COHERE *must* have an effort-like effect, while a violation of ALIGN need not have an effect, because it only puts constraints on the use of pronouns in *future* utterances.

3.2 Problems with original bidirectional OT

Although Blutner's bidirectional OT allows us to account for a number of conversational implicatures and can help to account for pronoun resolution insofar as it is able to explain the underlying principles of centering theory, there are serious problems with his analysis too. A major problem is that the analysis of *scalar implicatures* both *over* and *under*generates.[8]

3.2.1 Overgeneration

Blutner's (1998) original implementation of bidirectional OT overgenerates, because it predicts that whenever the semantic interpretation of *B*, $[[B]]$, entails the semantic interpretation of *C*, $[[C]]$, and the expressions *B* and *C* are equally complex, the assertion of *C* will have the scalar implicature that $[[B \wedge C]]$ is not true. The reason is that (on the assumption that worlds are reasonably equally distributed) for all $w \in [[C]] : \inf(w/[[B]]) > \inf(w/[[C]])$, which has the result that for these worlds the form–meaning pairs $\langle$'*B*',*w*$\rangle$ are *blocked* by the *S* principle. But this is obviously false. Although we normally conclude from assertion (3a) that the stronger (3b) is not true, we typically don't infer that (3c) is false from the assertion of (3b):

(3) a. John *believes* that Susan is sick.
b. John *knows* that Susan is sick.
c. John *regrets* that Susan is sick.

Suppose $D \models C$ and $C \models B$, and suppose that w is only true in D, v also in C, and u also in B. If we then assume that the worlds are equally distributed, Blutner's formalization gives us the following tableau:

inf(·/[[·]])	u	v	w
B	⇒1.4	1.4	1.4
C	*	⇒ 1	1
D	*	*	⇒ 0

Notice that in this tableau the values of inf(v/[[B]]) = inf(w/[[B]]) = 1, for example, because the semantic meaning of C leaves open only two alternative interpretations, [[C]] = {v,w}, and learning that it should be interpreted as v (or as w) gives us one bit of information. The double arrow indicates how the expressions should be interpreted according to Blutner's formalization. Because D is only true in w, it will be interpreted in that way. C, in turn, is interpreted as v, because (i) w can better be expressed as D, because inf(w/[[D]]) < inf(w/[[C]]), and (ii) v can better be expressed by 'C' than by 'B' because inf(v/[[C]]) < inf(v/[[B]]). From this tableau we can conclude that for any C and D, it holds that if the former entails the latter, we can infer from the assertion that D is the case that C is false. We can get rid of this false prediction, of course, by stipulating that ⟨know, believe⟩ forms a scale, but ⟨regret, know⟩ does not, that is, that C does not belong to the tableau. However, given the fact that the verbs "believe", "know" and "regret" are lexicalized to the same degree, it is not at all easy to explain this asymmetry.

3.2.2 Pragmatic scales

Blutner's analysis of Q-based implicatures, as any other analysis of scalars based on Grice's maxim of quantity, is also *not general enough*, because it cannot account for implicatures first discussed by Fauconnier (1975) and more extensively by Hirschberg (1985) that depend on *scales* where the meanings are logically independent and where the scalar behavior depends on the pragmatic context. For instance, it is of great value to have an autograph of a famous movie star. However, it doesn't count anymore to have one of Woodward when you already have one of Newman. Thus, we can conclude from (4b) that (4c) is false, but not the other way around:

(4) a. Did you get Paul Newman's autograph?
 b. I got Joanne Woodward's.
 c. I got Paul Newman's.

An analysis of this scalar implicature in terms of informativity would have to say that (4b) is more informative than (4c). That, however, seems to be unnatural. So, we must agree with Levinson (2000) that there are limits to the use of Bar-Hillel and Carnap's (1953) informativity function to account for scalar implicatures:

> Clearly, there are limits to the utility of such a characterization of informativity (e.g. rather a lot depends on what properties we are actually interested in). But, it is useful as a first approximation.
>
> (Levinson, 2000, p. 31)

The point of this section is that the limits of this approximation are rather disturbing. The main goal of this chapter, however, is to make clear that the claim with which Levinson continues the above quote is simply wrong:

> – and besides, it is just about the only measure of semantic information available.
>
> (Levinson, 2000, p. 31)

In the next section I will introduce measures of information, utility, or relevance that are much more appropriate to account for scalar implicatures than the "first approximation" used by Levinson and also by Blutner.

4 Maximizing utility

In this section I will first define a general decision theoretic notion of utility of propositions. I will then show that some specific measures that are found useful in accounting for linguistic phenomena turn out to be natural special cases of this general utility measure.

4.1 Decision theoretic utility

In Savage's (1954) decision theory, actions are taken to be primitives. If we assume that the utility of performing action *a* in world *w* is *U*(*a,w*), we can define the *expected utility* of action *a, EU*(*a*), with respect to probability function *P* as follows:

$$EU(a) = \sum_{w} P(w) \times U(a, w).$$

Let us now assume that our agent faces a *decision problem*, that is, she wonders which of the alternative actions in *A* she should choose. A decision problem of an agent can be modeled as a triple, ⟨*P, U, A*⟩, containing: (i) the agent's probability function, *P*; (ii) her utility function, *U*; and (iii) the alternative actions she considers, *A*. If she has to choose now, the agent simply should choose the action with the highest expected utility. But now suppose that she doesn't have to choose now, because she has the opportunity to first receive some useful information.

Before we can determine the utility of this new information, we first have to say how to determine the expected utility of an action conditional on learning this information. For each action a_i, its conditional expected utility with respect to new proposition *B*, $EU(a_i, B)$ is

$$EU(a_i, B) = \sum_w P(w/B) \times U(a_i, w)$$

When our agent learns proposition *B*, she will of course choose that action in *A* which maximizes the above value: $max_i EU(a_i, B)$. In terms of this notion we can determine the value, or *relevance*, of the assertion *B*. Referring to $a*$ as the action that has the highest expected utility according to the original decision problem, ⟨*P, U, A*⟩, that is, $max_i EU(a_i) = EU(a*)$, we can determine the *utility value* of the *assertion B*, $UV(B)$, as follows:[9]

$$UV(B) = max_i EU(a_i, B) - EU(a^*)$$

It seems reasonable to claim that in a cooperative dialogue one assertion or interpretation, *B*, is 'better' than another, *C*, just in case the utility value of the former is higher than the utility value of the latter, $UV(B) > UV(C)$.

4.2 Special cases

4.2.1 Topic value

The above way to determine the utility value of assertions is very general and follows from general and standard decision theoretic considerations. Now we focus our attention on two special cases, cases where only special kinds of actions are considered and where the utility functions are special too.

If only truth is at stake, a decision problem can be modeled by a partition of the logical space.[10,11] In Shannon's (1948) Information Theory, the **entropy** of partition *Q* w.r.t. probability function *P*, *E(Q)*, is defined as $\sum_{q\in Q} P(q)\times \inf(q)$, where inf(*q*) denotes the *informativity* of *q* that Blutner used already to implement his *H*-principle and is defined as $log_2 \frac{1}{P(q)}$. Thus, the entropy of Q is defined as follows:

$$E(Q) = \sum_{q\in Q} P(q) \times log_2 \frac{1}{P(q)}$$

This entropy $E(Q)$ measures the difficulty of the decision: the decision which element of Q is true is at its hardest when its elements are considered equally likely, and trivial in case one cell has probability 1. New information might *reduce* this *entropy*. Let us now denote the entropy of Q with respect to probability function P after B is learned by $E_B(Q)$:

$$E_B(Q) = \sum_{q \in Q} \underset{326}{P(q/B)} \times log_2 \frac{1}{P(q/B)}$$

Now we will equate the *reduction* of entropy, $E(Q)-E_B(Q)$, with the *Entropy value* of B with respect to decision problem Q and P, $EV_Q(B)$:

$$EV_Q(B) = E(Q) - E_B(Q)$$

Because learning B might flatten the distribution of the probabilities of the elements of Q, it should be clear that $EV_Q(B)$ might have a negative value. This won't happen when Q has a maximal entropy. The notion of entropy value gives rise to a linear order, $>$, on the usefulness of propositions, and we say that learning B is better than C in case $EV_Q(B) > EV_Q(C)$.

Suppose that partition Q has become relevant in a discourse either implicitly, or due to an explicit question asked by one of the participants, and that this question is very good in the sense that it has maximal entropy with respect to the relevant probability function. Now, there are two reasons why B could reduce Q's entropy more than C, that is, have a higher entropy value: either (i) because it *eliminates more cells* of the partition Q, or (ii) because it changes the probability distribution over the cells, that is, it makes some cells of Q that have a positive probability more probable than others. Assume that we ignore the latter possibility, that is, assume that when B is learned, each element of Q consistent with B has equal probability.[12] If we then quantify over probability functions, the above induced ordering relation comes down to the claim that B is better to learn than proposition C just in case B eliminates more cells of partition Q than C does:[13]

$$EV_Q(B) > EV_Q(C) \text{ iff } \{q \in Q: B \cap q \neq \emptyset\} \subset \{q \in Q: C \cap q \neq \emptyset\}$$

It is worth remarking that in this way we have reduced the ordering of propositions in terms of entropy reduction to the ordering between answers that Groenendijk and Stokhof (1984) have proposed.

Can we also think of reduction of entropy itself, that is, the entropy value of a proposition, $EV_Q(B)$, as a special case of the utility value of this proposition, $UV_Q(B)$ as discussed in the previous subsection? It turns out that we can (see van Rooy (2002) for proof) if we think of the alternative actions the decision maker considers in this case as probability distributions over the elements of *Q*.

4.2.2 Argumentative value

Ducrot (1973) argued that by making assertions we always want to argue for particular hypotheses, and analyzed linguistic expressions like *but* and *even* in terms of their argumentation orientation. More recently, Merin (1999) proposes the characterization of the contexts in which such expressions can be used appropriately in terms of their *argumentative value*, and proposes the implementation of this argumentative view on language use by means of probability theory. Suppose that an agent wants to argue for hypothesis h and that the relevant information state, that is, the common ground, is represented by probability function P. Notice that h is statistically dependent on proposition B iff learning B changes the probability of h, $P(h/B) \neq P(h)$. We might say that B is *positively relevant* with respect to h iff $P(h/B) > P(h)$. If $P(h/B) < P(h)$, B would be *negatively relevant*. Now we can define the *argumentative value* of proposition B with respect to hypothesis h, $AV_h(B)$, as follows:[14,15]

$$AV_h(B) \overset{def}{=} P(h/B) - P(h)$$

Assuming that an agent wants to argue for proposition h, we can order propositions linearly in terms of their argumentative value with respect to h. Thus, we can say that B is a better argument for h than C is iff $AV_h(B) > AV_h(C)$. Notice that this ordering relation might behave quite differently from one based on informativity: if B is consistent with h and C is not, $AV_h(B) > AV_h(C)$ even if $C \models B$.

Can we also think of the argumentative value of a proposition as a special case of its utility value? To do so we should resolve two questions: (i) what are the alternative actions? and (ii) what is the natural utility function involved? Notice that, just as in the previous case, only probabilities are at stake. So, it seems reasonable to assume that the decision problem (for a third participant) is now a choice of a probability measure. For worlds, such a probability measure comes down to a truth-value function. Because the speaker wants to be in a world where h is true, it's a truth-value function for h. The utility value can thus be defined as follows:

$$\begin{aligned} U(pr, w) &= 1, \text{ if } w \in h, \\ &= 0 \text{ otherwise} \end{aligned}$$

Now it is easy to see that the argumentative value of B with respect to 'goal' h is a special case of its utility value:

$$UV_h(B) = max_i(a_i, B) - EU(a^*)$$

$$
\begin{aligned}
&= max_i \textstyle\sum_{W} P(w/B) \times U\,(pr_i, w) - \textstyle\sum_{W} P(w) \times U(pr^*, w) \\
&= \textstyle\sum_{w\in h} P(w/B) - \textstyle\sum_{w\in h} P(w) \\
&= P(h/B) - P(h) \\
&= AV_h(B)
\end{aligned}
$$

5 Sperber and Wilson's relevance as utility

One of the central maxims of Gricean pragmatics is *Be Relevant*. Unfortunately, Grice stays rather vague about what he means with this maxim. Moreover, the constraint to be relevant seems to be just a qualitative condition, and not one that allows different interpretations to be compared with one another to see in how far they are relevant. Sperber and Wilson (1986/1995) have argued that interpretation is guided by the principle of relevance, stating that sentences should be interpreted as relevantly as possible:

> *The communicative principle of relevance*:
>
> Every utterance communicates a presumption of its own optimal relevance

For this principle to have some predictive force, we have to know what optimal relevance amounts to. According to Sperber and Wilson, the relevance of a proposition/interpretation depends on two factors: (i) the number of *contextual implications* that the interpretation gives rise to, and (ii) the *processing effort* needed to come to this interpretation (Sperber and Wilson, 1986/1995, p. 125).

> *Extent condition 1.* An assumption is relevant in a context to the extent that its contextual effects in that context are large.
>
> *Extent condition 2.* An assumption is relevant in a context to the extent the effort required to process it in that context is small.

When does one interpretation, *B*, give rise to more contextual implications than another, *C*? At first it seems that this is the case whenever *B* is *more informative* than *C*, that is, meaning that either *B* entails *C*, or that *B* rules out more worlds than *C* does.[16] The principle of relevance then seems to say that only in case *B* and *C* rule out equally many worlds, *B* is better than *C* if interpretation *B* is easier to 'grasp' than interpretation *C*. Although this seems to be Gazdar and Good's (1982), Merin's (1999) and Levinson's (2000) interpretation of Sperber and Wilson's notion of relevance, this can't be the reading they actually had in mind. For in that case it would be impossible to claim with Sperber and Wilson (1986/1995) that in the context of

(5a)–(5c), (6b) is not only more relevant than (6a), but also more relevant than (6c):

(5) a. People who are getting married should consult a doctor about possible hereditary risks to their children.
    b. Two people both of whom have thalassemia should be warned against having children.
    c. Susan has thalassemia.

(6) a. Susan, who has thalassemia, is getting married to Bill.
    b. Susan is getting married to Bill, who has thalassemia.
    c. Susan is getting married to Bill, who has thalassemia, and 1967 was a very good year for Bordeaux wine.

It is obvious that whether informativity is measured in terms of entailment, the number of worlds it eliminates, or the more abstract informativity function, 'inf', of Bar-Hillel and Carnap (1953), (6c) will come out as being more informative than (6b). However, when we think of increase of relevance as increase of utility value, in particular as increase of entropy value $UV\ (\cdot)$ as defined in the previous section, our analysis arguably makes better predictions. On the assumption that speakers are fully rational, and thus try to maximize their utility, we can assume that the speaker meant that interpretation of the sentence which has the highest utility value for the hearer. Thus, if sentence B with an underspecified meaning gives rise to a number of interpretations $B_i, \ldots, B_n$, the assumption gives rise to the hypothesis that the speaker meant that the interpretation with the highest utility will be chosen:

$$M(B) = max_i UV\ (B_i)$$

In case the speaker tries to maximize the entropy value, we have to assume that another agent faces a question that the speaker tries helping to solve. This seems a natural way to account for Sperber and Wilson's claim that (6b) is preferred to (6a). The reason is that in the above discourse two decision problems seem to be important that could be represented by the following two issues/questions (where the *Wh*-phrases range over Susan and Bill):

(7) a. Who should consult a doctor?
    b. Who should be warned against having children?

If we now assume that the number of contextual implications correlates positively with the number of eliminated cells of the partitions induced by the above questions, we predict that the number of contextual implications due to (6b) and (6c) is higher than that number due to (6a), and that (6c) doesn't give rise to more implications than (6b) does. Utterance (6a) resolves the first issue for Susan and Bill, while utterances (6b) and (6c) resolve also the second issue for both of these individuals. So, it seems not unreasonable to claim that one aspect of Sperber and Wilson's notion of relevance can be captured by our notion of utility.

However, Sperber and Wilson (1995) also claim that (6b) is more relevant than (6c), because the latter gives some extra *irrelevant* information which only costs *extra* interpretation *effort*. Fortunately, there is an easy way to capture this aspect of relevance too. Just say that in case the utility of B equals the utility of C, for example eliminates equally many cells of the salient partition, B is still more relevant than C in case the latter gives more information that is useless to solve the decision problem than the former (formally this means that relevance gives rise to a lexicographical ordering):

$$R(B) > R(C) \text{ iff (i) } UV(B) > UV(C), \text{ or}$$
$$\text{(ii) } UV(B) = UV(C) \text{ and } \inf(B) < \inf(C)$$

In case the utility value of proposition B is measured by the number of cells of the relevant partition that is eliminated, the ordering relation induced by relevance is almost the same as the ordering relation discussed by Groenendijk and Stokhof (1984) meant to capture the intuition when one answer is better than another. They claim that when B and C eliminate the same cells of a partition, B is still better than C in case C gives more information that is irrelevant to the question at hand, that is, when $C \subset B$.

I certainly don't want to suggest that Sperber and Wilson's notion of relevance is fully captured in the way described above. However, by making use of decision theory, a general theory of rationality that also applies to non-cooperative behavior, more aspects of their notion can be captured than just 'being an answer to a question':

> Achieving optimal relevance, then, is less demanding than obeying the Gricean maxims. In particular, it is possible to be optimally relevant without being 'as informative as required' by the current purposes of the exchange (Grice's first maxim of quantity): for instance by keeping secret something that it would be relevant to the audience to know. It seems to us to be a matter of common experience that the degree of co-operation described by Grice is not automatically expected of communicators.
>
> (Sperber and Wilson, 1986/1995, p. 162)

Indeed, when the goal is to make certain kinds of worlds true, or to argue for a particular hypothesis, maximal utility doesn't come down to being 'as informative as required', that is, to eliminate as many cells of the relevant partition as possible. In these cases the utility of a proposition is its argumentative value, and it might well be that to maximize this value one should not give as much information as possible: the probability of proposition h might be greater after learning just B, $P(h/B)$, than after learning the more informative proposition $B \wedge C$, $P(h/B \wedge C)$. In that case it is certainly more useful, though perhaps not very cooperative, to say only B.

So, I think that some aspect of Sperber and Wilson's notion of relevance can be captured by our very general decision theoretic notion of utility. In particular their notion of 'number of contextual implications' can be seen as correlating with being a 'good answer to a question'. The other side of their notion of relevance, the notion of 'processing effort' is obviously more dificult to formalize. However, at least some of the intuitions of Sperber and Wilson can be captured by assuming that in case two propositions, or two interpretations of a certain utterance, are equally useful, one is more relevant than another when the former gives less extra information than the latter.

Notice that this lexicographical analysis (above) allows us to account for some examples that typically involve stereotypical interpretations. A sentence like (8a) is typically interpreted as (8b) because it is the most probable meaning:

(8) a. John said 'Hello' to the secretary.
b. John said 'Hello' to the *female* secretary.

We can account for an example like this, as for other so-called *I*-inferences discussed in Section 3, if we assume that its stereotypical interpretation and its alternative(s) are equally useful. In that case we predict that the most probable meaning is the most relevant one, giving rise to the stereotypical interpretation.

I don't believe, however, that by taking utility and effort into account in a lexigraphical way as suggested above I can analyze successfully all the kinds of examples Sperber and Wilson's notion of relevance is meant to take care of: I predict that a more stereotypical interpretation of an utterance is preferred only if none of the other interpretations is more useful, whereas they seem to suggest that a stereotypical interpretation can be the most relevant one although there might be other interpretations that, after all the processing is done, turn out to have (in my terms) a higher utility value.[17]

> …the order in which hypotheses are tested affects their relevance. As a result, the principle of relevance does not generally warrant the selection of more than one interpretation for a single ostensive stimulus.
>
> …Consider the following utterance, for instance:
>
> (65) George has a big cat.

> In an ordinary situation, the first interpretation of (65) to occur to the hearer will be that George has a big *domestic* cat. … the first interpretation consistent with the principle of relevance was the best hypothesis. All other interpretations would manifestly falsify … the presumption of relevance.
>
> (Sperber and Wilson, 1986/1995, pp. 167–8)

Although my lexicographical analysis of relevance doesn't seem to be fully adequate/sufficient to capture the effects of effort, we will see that by thinking of my notion of 'maximizing relevance' as only one of the two guiding principles of bidirectional OT, some other effects of 'minimizing effort' can be captured in this more general framework.

6 Stalnakerian constraints and Gricean maxims

In Section 3 of this chapter I have shown how Beaver's (to appear) OT constraints used to capture centering theory could be motivated by reducing them to Blutner's general informativity function. In this section I want to do something very similar with respect to other constraints used in OT to account for semantic/pragmatic phenomena. In particular, I want to discuss to what extend Stalnaker's assertability conditions and Grice's conversational maxims can be motivated by the general presumption of optimal relevance/utility in combination with Blutner's bidirectional OT. Grice's maxim of quantity, and the implicatures it is usually said to account for, will be our main concern. Because both Stalnaker and Grice assume that participants of a conversation behave cooperatively, this section will deal almost exclusively with utility value instantiated as entropy reduction.

6.1 Stalnaker's assertion conditions

In his very influential article 'Assertion', Stalnaker (1978) states three principles that have come to be known as Stalnaker's assertion conditions that he claims "can be defended as essential conditions of rational communication". Let's see to what extent these three principles can be based upon our decision theoretic approach. I will discuss them in reverse order.

6.1.1 Avoid ambiguity

Stalnaker's third principle basically says that speakers should *avoid ambiguity*. Can this principle be motivated from our decision theoretic point of view? I think we can. First, note that according to our analysis, a sentence can be truly ambiguous only if there are at least two interpretations of this sentence that are optimally relevant. Now suppose our hearer faces a decision problem and hears a truly ambiguous sentence. In that case it might be that according to one interpretation the agent is advised to do one action, for example *a*, while according to the other interpretation he is advised to do action *b*. This has the result that the hearer doesn't know what to do, and, worse, might choose the wrong action. This is certainly something we don't want a cooperative speaker to be responsible for, and thus we shouldn't allow her to use a truly ambiguous sentence.[18]

6.1.2 Presupposition

Stalnaker's second condition advises the speaker to use only sentences that express a proposition in each world of the context, which means that (certain kinds of) its linguistic presuppositions have to already be common ground. It appears to make little sense to make this principle a hard constraint: although the verb *know* is normally assumed to trigger a factive presupposition, it is not really problematic to use a sentence like *John knows that Mary is coming* even though Mary's coming is not yet common knowledge. Although such examples seem to violate the principle, it is standardly assumed with Lewis (1979) that the constraint can be rescued by assuming that in these cases we first *accommodate* the context such that the principle holds after all. Be that as it may, it still seems bad conversational practice to change contexts by means of presupposition accommodation. Moreover, some presuppositions seems to be accommodated more easily than others. In fact, in their use of OT to account for semantic/pragmatic phenomena, Zeevat (1999, 2000) and Aloni (2001) propose a violable constraint to ban presupposition accommodation. Can we give an explanation for why this constraint makes sense?

The explanation cannot be straightforward by using our analysis of relevance: presupposition accommodation enriches the context with new information and we saw that new (consistent) information can never have a negative utility. To explain why it is better conversational practice to enrich the context by asserting it than by presupposing it, we have to distinguish the ways in which presupposition and assertion are allowed to change the context. In a rich and very stimulating article, Merin (1999) proposes that (argumentative) relevance helps here: he claims that presupposition *B* is allowed to be informative with respect to the context, but that this new information should not have a positive relevance. I find this proposal very intuitive, but I don't think it can be a hard constraint: though perhaps not very polite, I find it sometimes a useful strategy to influence people *indirectly* by means of presupposition. Moreover, it is unclear to me how the presumption of optimal relevance can explain Merin's proposal.

Although I am not able (yet?) to explain the ban on accommodation by a presumption of optimal relevance,[19] a closely related principle proposed by Van der Sandt (1992) that prefers binding to accommodation seems to have a natural relevance-theoretical explanation. The principle says that if new information is accommodated to the context, it is better to *bind* this new information to already existing *discourse referents* of the context than to introduce new such referents. The 'motivation' for this principle is based on the fact, to be discussed in Section 6.2, that in special cases maximizing utility comes down to maximizing informativity. If the context already contains the information that a certain (underspecified) individual has property P, and it is presupposed (by a presupposition trigger like *too*) that somebody has property Q, it is more informative to assume that it is the same individual having property P and Q than to assume that the properties are distributed over (possibly) different individuals ($\inf(\exists x[Px \wedge Qx]) > \inf(\exists xPx \wedge \exists yQy)$). In fact, this explanation is the natural analogue to Levinson's (2000, p. 273) explanation of why coreference is preferred to disjoint reference. Unfortunately, however, I am not at all convinced of this explanation of the preference for coreference. I find explanations in terms of maximizing *coherence* between clauses proposed by proponents of centering theory, and by authors like Hobbs (1979) and Asher and Lascarides (1998) much more natural. Remarkably enough, as we saw in Section 3.1, the centering theoretical explanation for the preference for coreference can be motivated by the opposite assumption that expressions should be interpreted in the *least surprising* way: the interpretation selected is the one for which the (conditional) informativity is *lowest*. As we saw in Section 5, this follows from the presumption of optimal relevance (from the hearer's point of view) only if we make the counter-intuitive assumption that the utility of the resulting interpretation of the sentence in which the pronoun occurs is independent of the choice of reference of the pronoun.

6.1.3 *Be consistent!*

Stalnaker's first assertion conditions demands two things: (i) to be consistent, and (ii) to be informative. To motivate (i), we have to see why inconsistency is bad.

Suppose B is inconsistent with $W(P) = \{w \in W : P(w) > 0\}$. Now there are two possible explanations. According to the standard way we say that $P(C/B)$ is undefined in case $B \cap W(P) = \emptyset$. It seems reasonable to stipulate that in that case $UV(B)$ is undefined as well, 'explaining' why learning information inconsistent with the context is bad. But we are obviously able to learn new information that is blatantly inconsistent with what we believed before. Can we give a decision theoretic motivation for why speakers should be consistent with what is commonly assumed even if we take this fact seriously? Suppose we allow $P(\cdot/B)$ to be defined even though B is inconsistent with $W(P)$, but that the result will be that P is *revised* by new information B, resulting in probability function $P^*_B(\cdot)$,[20] with the effect that $W(P) \cap W(P^*_B) = \emptyset$. The problem of revision with inconsistent information, however, is that it is normally not clear what the best way to do so is: there are typically more alternative ways to revise one's belief state that are equally optimal. In our case this means that there are typically several P^i_Bs that count as optimal revisions of P by B. Because the agent can't choose between them, he doesn't. He either feels 'ambiguous' about which belief state he is in, and the motivation given in the previous subsection applies here as well or, alternatively (but less naturally), we might represent his belief state as a linear combination of the optimal probability functions after revision. According to this latter possibility, many more worlds will be consistent with the new probability function than with the old one. This has the result that there might be many more actions than the ones considered before that could be optimal in (at least) one of the worlds consistent with what is believed, which means that the *risk* that our agent will choose the wrong action has *increased*.

6.1.4 *Be informative!*

The second part of Stalnaker's first assertion condition demands that new information has to be *informative* with respect to what is commonly assumed, that is, the context represented by our probability function P. Suppose now that our utterance has B as a relevant interpretation and thus has a strictly positive utility value: $UV(B) > 0$. Then it is easy to see that this interpretation must also be informative, that is, incompatible with at least some worlds in $W(P)$.

If $UV(B) > 0$, it has to be the case that $max_a \Sigma_w P(w/B) \times U(a, w) > \Sigma_w P(w) \times U(a^*, w)$. This, however, can be the case only if either learning B has the result that an action different from a^* has the highest expected utility afterwards and thus will be chosen, or the preferred action remains the same, but the expected utility of this action is higher after learning B than before. But either one of those can happen only in case B at least eliminates some worlds in $W(P)$ and thus is informative. Because the entropy value and argumentative value are both special cases of utility value, we have shown that old 'news' can never be useful. The other way around, however, doesn't follow: A proposition can be informative with respect to probability function P without being relevant.

6.2 Maximal informativity: the *I*-principle

In the previous subsection we saw that a necessary condition for a proposition to be relevant, or useful, is to be informative. In case an utterance allows for more than one interpretation, our analysis predicts that the preferred one should at least be informative. According to Atlas and Levinson's (1981) and Levinson's (2000) *I*-principle and Horn's (1984) *R*-principle something more is demanded: the preferred interpretation is the one which is *maximally informative*.[21] Although, as we saw in Section 3, there are good reasons not to assume this principle in its full force, it seems to make correct predictions for a certain range of phenomena. In this section I show that in certain special circumstances usefulness reduces to informativity.

6.2.1 *Entropy value*

First, note that it is obvious that in case B *eliminates* more *cells* of the relevant partition than C does and cells are taken to be as fine-grained as worlds, eliminating more cells means eliminating more worlds. On the extra assumption that the cells of the partition are equally likely, it also means that B has in that case a higher 'inf' value.

Second, this result generalizes quite straightforwardly when relevance is measured in terms of *reduction of entropy*. If W is the set of all worlds, the entropy value of proposition B, $EV_W(B)$, is then $E(W) - E_B(W)$. It is obviously the case that this value gets higher when the entropy of W conditional on B gets lower. Thus, if we can show that $E_B(W) < E_C(W)$ iff $\inf(B) > \inf(C)$, we show that in these special cases maximizing entropy reduction comes down to maximizing informativity. As shown in van Rooy (2002), this can indeed be done in case the probabilities are equally distributed over the worlds. To illustrate, notice first that for every world w it holds that $w \in B$ or $w \notin B$, so that we can equate $E_B(W)$ with $\Sigma_{w \in A} P(w/B) \times - log_2 P(w/B)$. Suppose now that we have eight worlds, and that $P(B) = 1/4$. Then B is true in two of the eight worlds, and thus $E_B(W) = 2 \times (\frac{1/8}{1/4} \times - log_2 \frac{1/8}{1/4}) = 2 \times (1/2 \times - log_2 \frac{1}{2}) = 2 \times 1/2 = 1$. Now suppose that $P(C) = 1/2$, and thus that C is true in four of the eight worlds. In that case it holds that $E_C(W) = 4 \times (\frac{1/8}{1/2} \times - log_2 \frac{1/8}{1/2}) = 4 \times (1/4 \times 2)_1 = 2$.Because $E_B(W) < E_C(W)$ it also is the case that $EV_W(B) > EV_W(C)$. We can conclude that in these special circumstances the relevance of proposition B is higher in case its probability is lower, which holds exactly when its informativity value, $\inf(B)$, is higher. Thus, in these circumstances reduction of entropy is monotone increasing with respect to informativity, and maximization of the one comes down to maximization of the other.

6.2.2 Argumentative value

Finally, we can show that in special cases the *argumentative value* of a proposition is also monotone increasing with respect to its 'inf' value. In Section 4.2.2 we argued that proposition B has a positive argumentative value with respect to h, that is, $AV_h(B) > 0$, just in case $P(h/B) > P(h)$. Notice that $P(h/B) > P(h)$ iff $P(h/B)/P(h) > 1$ iff $P(B/h)/P(B) > 1$. In fact, the measure $P(\cdot/h)/P(\cdot)$ is continuously monotone increasing with respect to our $AV_h(\cdot)$, meaning that if the one gets higher (lower), the other gets higher (lower) too. Notice that when $h \vDash B$, $P(B/h)/P(B) = \frac{1}{P(B)}$. The function $\frac{1}{P(\cdot)}$, in turn, is continuously monotone increasing with respect to Bar-Hillel and Carnap's (1953) informativity function, because $\inf(\cdot) = log_{\frac{1}{P(\cdot)}}$. Thus, if h entails the arguments given, the measure $P(\cdot/h)/P(\cdot)$ is continuously monotone increasing with respect to $\inf(\cdot)$. But this means that in these cases also our $AV_h(\cdot)$ is continuously monotone increasing with respect to $\inf(\cdot)$. We can conclude that in special circumstances the requirement to select the maximally relevant interpretation of a sentence comes down to selecting its most informative interpretation.

6.2.3 Sufficiently informative

In this section I have interpreted the *I*-principle as the demand to interpret the sentence in the most informative way in the sense of Bar-Hillel and Carnap's (1953) informativity function 'inf'. Although Horn, and especially Levinson, make use of the *I*- (or *R*-)principle under this interpretation, their explicit statement of the principle actually demands only that the most informative interpretation "sufficient to achieve your communicational ends" (Levinson, 2000, p. 114) be taken. And indeed, under this interpretation the *I*-principle is close to what Grice's (1989) second maxim of quantity asks for. Notice that in case relevance is measured in terms of entropy value, we might say that informativity is measured with respect to the goals/topics the discourse participants are interested in. Before we discuss such an interpretation of Grice's maxim, however, it is useful to first discuss his maxim of manner, and see to what extent our analysis can capture it.

6.3 Manner

Grice's maxim of manner asks the speaker to be *perspicuous*, which by itself gives rise to the following four (sub)maxims:

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

As Grice notes himself, the maxim of manner is rather different from the others because it relates "not (like the previous categories) to what is said but, rather, to HOW what is said is to be said". Still, the first two submaxims can, I believe, be motivated by our general decision theoretical approach in the same way as I motivated Stalnaker's third assertion condition. The other two submaxims seem to be very close to Zipf's principle of *minimizing effort*, a principle that was already captured adequately, or so we argued, by Blutner's interpretation of the *S*-principle in his bidirectional OT.

This, then, suggests a way of combining the presumption of optimal relevance/utility with bidirectional OT: Blutner's *S*-principle stays as it is, capturing Grice's last two submaxims of manner and part of Zipf's minimization of effort, but his ordering on interpretations used in the *H*-principle should be induced (at least in a number of cases) by the above discussed notion of relevance.[22] If we do that, we are ready to see to what extent we can account for the effects of Grice's maxim of quantity.

6.4 Quantity and *Q*-implicatures

6.4.1 The maxims and their interpretations

Grice's maxim of quantity talks about the quantity of information to be provided, and thus seems most closely related with our quantitative analysis of relevance. Quantity comes with the following two maxims:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

As we have seen in Section 1, the first maxim of quantity was interpreted by Horn, Gazdar and others as meaning something like 'say as much as you can', and this maxim was taken to be responsible for many so-called *generalized* conversational implicatures: the *scalar* and the *clausal* ones. However, as noted by Gazdar (1979), it is not straightforward to interpret and/or formalize the maxim in its full generality:

> To formalize this maxim as it stands, that is in its full generality, we would have to (*a*) be able to quantify over informativeness, and (*b*) have some function which when applied to a conversation and a point within it would yield as its value the level of informativeness required.
>
> (Gazdar, 1979, p. 49)

It seems that our analysis of topic-dependent relevance, that is, entropy value, provides exactly what Gazdar asked for. According to our treatment, *B* can only have a higher entropy value than *C* in case it is more informative. Moreover, our topic-dependent analysis of relevance also says in what sense a sentence can be more informative than required: In case proposition *B* resolves the decision problem, any stronger proposition *C* will resolve the decision problem too. In that case, however, *C* will give extra, irrelevant, information and, according to our analysis of relevance, interpretation *B* is then preferred to interpretation *C*.

So, how does the presumption of optimal relevance relate to Grice's quantity maxims? First, it predicts that an agent might give information that is maximally relevant without being 'as informative as required'. In case you want to argue for hypothesis *h*, or make true a world where *h* holds, it might be more useful to say less than is required. In those cases, of course, Grice's cooperative principle is not at work, so the deviation should not come as a big surprise.

However, when we limit ourselves to utility as entropy reduction, the proposal to ask for the hearer to interpret the utterance as maximally relevant seems to have a straightforward connection with Grice's quantity maxims. But first we have to make clear how we understand these maxims.

According to the standard reading of the first maxim of quantity, as interpreted by Horn, Gazdar, Levinson and others, it says that the speaker could not make an alternative claim relevant to the conversation with a *stronger/more specific* conventional/semantic meaning. In this reading this maxim is responsible for the standard treatment of scalar and clausal implicatures.

The second maxim of quantity is normally (e.g., Horn, Levinson) taken to mean the opposite: it allows the speaker to use a sentence with a very *weak/general* conventional/semantic meaning if she can rely on the hearer to interpret the sentence in the intended stronger/more specific way because that is what the purpose of the exchange requires.

The proposal to ask the hearer to interpret the utterance as relevantly as possible seems in accordance with Grice's second submaxim of quantity, but in contradiction with the first one.

However, there might be another way to interpret Grice.[23] Suppose that Grice, in stating his maxims, already took the hearer's perspective into account. In that case, Grice's first maxim of quantity says something very close to our demand to choose that interpretation of a sentence which has the highest entropy value, while the second maxim can then be interpreted as saying that in case two interpretations of a sentence have an equally high entropy value (for instance, if both completely resolve the issue), the less informative one is preferred. On this reading of Grice, quantity reduces to our lexicographical definition of relevance repeated below:

$$R(A) > R(B) \text{ iff (i) } UV(A) > UV(B), \text{ or} \\ \text{(ii) } UV(A) = UV(B) \text{ and } \inf(A) < \inf(B)$$

Notice that it is the first submaxim that is standardly used to derive scalar and clausal implicatures. However, as we saw in Section 3, this maxim is also responsible for the overgeneration: for instance, it is not clear how we can rule out a scale like ⟨*Regret, Know*⟩ other than by stipulation. As observed by Groenendijk and Stokhof (1984), and earlier by Atlas and Levinson (1981), it is also responsible for the false prediction that answers to *Who*-questions are typically *not* interpreted as being exhaustive, for otherwise this would have been done explicitly. Perhaps, despite its overwhelming use in (neo-)Gricean pragmatics, we can, and thus should, do without Grice's first maxim of quantity once we have our principle of optimal relevance together with bidirectional OT. In the remainder of this chapter we will see how far we can pursue this line of thought.

6.4.2 'Exactly' interpretation of numerals

A first example. We have to account for the fact that in most contexts number terms get an 'exactly' interpretation. At the same time (cf. Kempson, 1986; Kadmon, 1987; Zeevat, 1994; van Kuppevelt, 1996), the analysis should also explain why the sentence

(9) John has three children.

does not get this ‘exactly’ interpretation when given as an answer to the question:

(10) Does John have three children?

Let us assume with neo-Griceans that number-terms semantically get an ‘at least’ meaning. In that case the second maxim of quantity, and our maxim of optimal relevance, seem to do the trick: When the question is

(11) How many children does John have?.

the question gives rise to the partition $\{\lambda v[\text{John has exactly } n \text{ children}]: n \in N\}$, where each cell contains only worlds where it is true that John has *exactly n* children. Thus the exact number of children that John has is relevant, and we should look for the most informative reading of (9). But what is this most informative reading? Assuming that one reading is more informative than another if it eliminates more cells of the partition, it should be a reading of the form λv[John has exactly n children] that is compatible with the semantic meaning of (9): λv[John has at least n children]. Intuitively, this most informative reading should be the one saying that John has exactly three children. Unfortunately, informativity by itself cannot enforce this reading: The reading ‘exactly 3’ is not the only one compatible with (9) when numerals have an ‘at least’ meaning; a reading like ‘exactly 4’ is so too. Why should (9) not be interpreted as an exhaustive answer incompatible with John’s having exactly three children? The question seems silly, but this is only so because we take the answer to be so obvious: because for the other cells we use other numbers. Thus, alternative expressions should come into the picture after all. And with the alternative expressions, then also Blutner’s bidirectional OT.

If we then assume that the probabilities are equally distributed over the worlds, that it is already assumed that John has children, but not more than four, a bidirectional formalization in terms of relevance gives rise to the following tableau, with the desired outcome:

$EV_{(11)}([[\cdot]])$	1	2	3	4
‘1’	⇒0	0	0	0
‘2’	*	⇒ 0.4	0.4	0.4
‘3’	*	*	⇒ 1	1
‘4’	*	*	*	⇒ 2

Notice that '3' doesn't mean four because this meaning is *blocked*: there is another expression for which four is a better meaning, that is, a meaning with a higher relevance.[24]

6.4.3 Cancellation

Notice that this much could already have been done by Blutner's (1998) ordering relation between meanings in terms of (conditional) informativity. But we also saw that analysis was both (i) too general, and (ii) not general enough. Let us first discuss the cases that Blutner's ordering in terms of informativity could not account for.

First, when the question answered by (9) is not (11) but (10) instead, the answer intuitively does not rule out that John has four children. When the meanings/worlds are ordered by conditional informativity inf($m/[\![\cdot]\!]$), however, this is what is predicted: informativity alone doesn't care about relevance. To make correct predictions, Blutner (1998), following standard analyses of conversational implicatures, would have to allow for implicatures that can be *canceled* for reasons of relevance. When the ordering depends on relevance, on the other hand, things are different. In that case both answer '3', that is, (9), and answer '4' would have a relevance of one in worlds where John has three or four children. Because '3' and '4' are equally complex, bidirectional OT predicts that (9) now does not give rise to the inference that John doesn't have more than three children. Thus, no cancellation is needed, just like Kadmon (1987), Zeevat (1994) and van Kuppevelt (1996) propose.

6.4.4 Exhaustivity

We have seen above that bidirectional OT predicts that answers involving numerical terms are interpreted *exhaustively*. Groenendijk and Stokhof (1984) make use of an explicit exhaustivity operator to account for this. But their operator accounts not only for standard *scalar* inferences, but also for the intuition that when (12b) is given as an answer to (12a), the answer is interpreted as meaning that *only* John went to the party:

(12) a. Who went to the party?
 b. John went to the party.

As Groenendijk and Stokhof (1984) note themselves, this is certainly not an inference following from Grice's first maxim of quantity. That maxim would rather predict that the answer should *not* be interpreted exhaustively. However, the inference *does* follow in bidirectional OT from the assumption that answers should be interpreted maximally relevant. Suppose that only *a* and *b* are the relevant persons for question (12a). In that case the bidirectional tableau looks as follows:

$EV_Q([[\cdot]])$	$\emptyset$	*a*	*b*	*ab*
‘Nobody’	⇒ 2	*	*	*
‘a’	*	⇒ 1	*	1
‘b’	*	*	⇒ 1	1
‘ab’	*	*	*	⇒ 2
‘not a’	1	*	1	*
‘not b’	1	1	*	*
‘not a and not b’	2	*	*	*

Notice that in this tableau complexity plays a crucial rule. Meaning *b*, for example, is expressed by ‘*b*’ and not by ‘not *a*’ because the former is less complex than the latter. From this table we can conclude that in this example (12b) is predicted to mean that John was the *only* one who went to the party. This seems perfect. Still, as we will see in Section 7, the analysis of exhaustivity can’t be so straightforward anymore once we look at examples just a little bit more complicated than the one discussed here. But before we come to that, let us first discuss some cases that can’t be handled straightforwardly by making use of Groenendijk and Stokhof’s explicit exhaustivity operator, but that are unproblematic on our account.

6.4.5 Mention some

To account for the intuition that (12b) is treated as an *exhaustive* answer to question (12a), we have assumed that the decision problem is which answer to question (12a) is true, and that the question itself gave rise to a partition. But we might give up both of these assumptions. First, we might assume that the question is not represented as a partition, but treated as a mention-some question where its answers might overlap. If the worlds are the same as in the above example, and if it is assumed that at least one of {*a,b*} went to the party, the question can be represented as {{*a,ab*},{*b, ab*}}. Second, we might propose that the question is still represented as a partition, but that the decision problem is such that one action is best in world *a*, the other in world *b*, but both are equally good in world *ab*. Whether we now determine the relevance of answers with respect to the non-partitional question in the first case, or with respect to the decision problem in the second, the entropy value will be the same. In both cases the possible answers give rise to the following tableau:

$EV_Q([[\cdot]])$	*a*	*b*	*ab*
'a'	⇒ 1	*	⇒ 1
'b'	*	⇒ 1	⇒ 1
'ab'	*	*	1
'not a'	*	1	*
'not b'	1	*	*

Notice that in this case (i) the answers 'a' and 'b' are not interpreted exhaustively, and (ii) it is predicted that answer 'ab' will not be given, because there is no need to specify this world separately by the use of a more costly expression. It seems to me that both predictions are born out by the facts.

6.4.6 Pragmatic scales

We have seen in Section 2 that informativity (alone) cannot account for the fact that in the context of question (4a), repeated here as (13a), we conclude from (13b) that (13c) is true, but we don't infer (13e) from (13d):

(13) a. Did you get Paul Newman's autograph?
b. I got Joanne Woodward's.
c. I didn't get Paul Newman's.
d. Yes/I got Paul Newman's.
e. I didn't get Joanne Woodward's.

An analysis in terms of relevance can do much better, but now we have to use Merin's (1999) notion. This seems reasonable in this case: the answerer wants to convince the questioner to accept that we are in a world where she has an autograph of somebody with a high prestige, and, if possible, an autograph with a higher prestige than the questioner himself. Let us assume that the questioner does not yet know that the answerer got an autograph of a famous movie star in the first place, that having an autograph of such a person is of great value, but that it doesn't count anymore to have one of Woodward when you already (or also) have one of Newman. In that case we get something like the following tableau (where the numbers might be different, but the ordinal relations between the numbers remain the same):

$AV_h([\![\cdot]\!])$	$\neg N \wedge \neg W$	$\neg N \wedge W$	$N \wedge \neg W$	$N \wedge W$
'No'	$\Rightarrow 0$	0	0	0
'Woodward'	*	$\Rightarrow 0.7$	*	0.7
'Woodward and not Newman'	*	0.7	*	*
'Yes'	*	*	$\Rightarrow 1$	$\Rightarrow 1$
'not Woodward and Newman'	*	*	1	*

Notice that in this case the answers where both persons are mentioned are ruled out for reasons of speaker effort, and that relevance does the rest.

6.4.7 Limiting overgeneration

In this section we have seen that by replacing the informativity ordering relation on meanings by one of relevance, we can account for more scalar implicatures than before. But this analysis overgenerates neither as much as the ordering relation that Blutner proposed, nor as much as the standard neo-Gricean (e.g., Horn, Levinson) treatment of scalar implicatures in terms of Grice's first maxim of quantity. In Section 3 we saw that ordering by informativity wrongly predicts that if B follows from C, the assertion 'B' always gives rise to the implicature that C is false. This prediction doesn't follow anymore once we order meanings in terms of relevance. The reason is that although B might follow from C, this doesn't necessarily mean that in the $B \wedge \neg C$-worlds assertion 'C' has a higher relevance with respect to the question under discussion than assertion 'B'. In fact, if the extra information that C asserts on top of B is *irrelevant* to the topic of the conversation, it is predicted that the relevance of C in those worlds is *lower* than the relevance of B. For instance, in case the question is how sure John is that Susan is sick, it is predicted that in every world where John knows that Susan is sick, (14a) has a higher relevance than (14b):

(14) a. John *knows* that Susan is sick.
b. John *regrets* that Susan is sick.

This gives rise to the correct prediction that in the context of such a question (14a) does not give rise to the inference that (14b) is false. I conclude that in combination with bidirectional OT, the assumption of optimal relevance predicts better with respect to scalar implicatures than Grice's first maxim of quantity under its standard reading.

Green (1995) has argued that the wrong prediction of neo-Griceans is due to a wrong reading of Grice's first maxim of quantity. Neo-Griceans have standardly assumed that Quantity 1 means that the speaker is making the strongest statement she is able to make on the matter at hand (i.e., saying as much as she can). Green argued that Grice only requires, however, that the speaker makes a contribution which is (at least) as informative as is required, that is, *informationally sufficient*. But if that is so, and if we also assume that Quantity 2 means that the speaker should not say something stronger than is required, it seems that Grice himself already correctly predicts that in the context of the question described above, (14a) doesn't give rise to the implicature that (14b) is false. I think Green gives a new, interesting, and empirically more adequate, interpretation of Grice's maxim. Be that as it may, to formalize this reading of Grice, we have to say what it means to be as informative as *required*. To account for that, however, it seems we still need a notion of relevance. The purpose of this subsection, however, was to argue that once we have a notion of (optimal) relevance, in combination with bidirectional OT, we do not need the Gricean maxim of quantity anymore.

7 Maximization of relevance as exhaustification

In the previous section we have seen how our use of relevance in bidirectional OT explains why an answer like *John went to the party* to the question *Who went to the party?* is typically interpreted exhaustively when the interrogative sentence should be interpreted as a mention-all question. But I have already noted that things are not as straightforward as they seem. There are (at least) two reasons for this: (i) we limited ourselves to the simple case where only a few individuals were taken to be relevant, (ii) we considered only how to encode the *cells* of a partition and have not taken *partial answers* into account. With respect to the second problem, we have not discussed yet the perhaps most obvious problem for the standard analysis of scalar implicatures: the fact that from the answer '*a* or *b*' to the question *Who is coming?* it is wrongly predicted that neither *a* nor *b* will come. The reason for this false prediction is that both the answer *a* and the answer *b* would entail the answer actually given, and thus, by the standard reading of Quantity 1, are ruled out.[25] Our analysis does not generate this problem, but gives rise to another one: how should we interpret '*a* or *b*' in the first place, and how can we explain that such a disjunctive answer normally gives rise to an *exclusive* reading? One might try to extend the bidirectional analysis by taking more alternative expressions into account, and also more meanings than just the cells of the partition. As it turns out, this is not a trivial enterprice. Instead of getting involved in this enterprise, let me discuss another problem of our approach which suggests a somewhat different line of attack.

In Sections 2 and 3, I have shown the potential of bidirectional OT when meanings are ordered in terms of Blutner's conditional informativity function. After that I have argued that with this way of ordering meanings we encounter difficulties in accounting for certain examples and I have shown that bidirectional OT makes better predictions if we assume, with Sperber and Wilson, that sentences are interpreted as relevantly as possible. To account for that we assumed that meanings are ordered in terms of our decision theoretic notion of utility. Although we saw in the previous section that by making use of relevance/utility in bidirectional OT we can account for many *Q*-implicatures, it should be clear that such an analysis is not really suited to account for *I*-implicatures. To account for these latter kinds of implicatures we had to assume that the information given is *irrelevant*. Our discussion of why stereotypical interpretations, and in particular coreferential interpretations of pronouns, are preferred suggested, however, that this assumption is implausible and that our lexicographical analysis of relevance isn't quite satisfactory. Thus, it seems that if we want to account for implicatures in terms of a single general function, we either have to use something like the conditional *informativity* function as used by Blutner, or the assumption that we interpret things as *relevantly* as possible and account for that in decision theoretical terms. If we choose the first option, we can account for *I*-implicatures to stereotypical interpretations, but we can't account for *Q*-implicatures. If we go for the second option, however, it is rather the *I*-implicatures that we cannot account for adequately anymore. So it seems that our search for a single general principle in terms of which all kinds of implicatures can be handled ended unsuccessfully. In this final main section, however, I want to suggest that prospects are not that dim.

The new idea is to shift once again to another reading of Grice's maxims. First, we followed Blutner (1998) in taking his interpretation principle based on the conditional informativity function as an implementation of Grice's *first* submaxim of quantity as understood by Horn, Levinson and others: Say as much as you can! Afterwards, we have used utility in accordance with Sperber and Wilson's principle to *interpret* sentences as maximally relevant, which can be based on Grice's *second* submaxim of quantity: Don't say more than you must! But perhaps we should make use of utility not from the hearer's, but rather from the *speaker's* point of view. In that case it seems natural to use utility to interpret Grice's first maxim of quantity, so that it reads: Speak as relevantly as you can! From our earlier discussion it seemed that if we want to account for *Q*-implicatures in terms of Grice's first maxim of quantity, we have to make crucial use of *alternative expressions*. This use of alternative expressions, however, was seen to be dangerous: without limitations the analysis would overgenerate enormously. In this final section I would like to suggest that by adopting an *exhaustivity operator* – in fact by changing Groenendijk and Stokhof's (1984) context-independent exhaustivity operator into one that is based on a relevance-ordering – we can actually account for both *I*-implicatures and *Q*-implicatures with just one operation.

Groenendijk and Stokhof (1984) propose to account for the intuition that the answer *Peter comes* to the question *Who comes?* should normally be read exhaustively by introducing an explicit exhaustivity operator that is applied to answers and the abstracts (predicates) underlying the questions to derive the exhaustive interpretation. Although their exhaustivity operator is very appealing and predicts correctly when assertions are given as answers to so-called *mention-all* questions, it also faces some crucial problems. First, it gives the wrong result if applied to answers given to *mention-some* questions. Second, it cannot account for Hirschberg's examples of *scalar* readings. To solve both of these problems, the following exhaustivity operator can be defined (see van Rooy and Schulz, 2003) which is dependent on a relevance-ordering '>':

$$[[exh]] = \lambda T \lambda P.\{w \in W | \; P(w) \in T(w) \wedge \neg\exists t \in T(w): \lambda v[P(w) \subseteq P(v)] > \lambda v[t \subseteq P(v)]\}$$

This operator takes a term-answer T and a question-predicate P and turns it into a proposition. Described informally, it does the following: in each world, T denotes a generalized quantifier, that is, gives a set of possible extensions for P. *exh* takes all these possibilities $t \in T(w)$ and compares the utility value of the propositions $\lambda v[t \subseteq \mathrm{P}(v)]$. P can only be one of these possibilities that are minimal values in this order. This exhaustivity operator can be thought of as a generalization of Groenendijk and Stokhof's exhaustivity operator. The two operators give rise to (almost) identical results in case the relevance ordering '>' reduces to entailment, or the subset relation '⊆'. As a consequence, our operator accounts for many of the implicatures traditionally accounted for in terms of Grice's maxim of quantity. Just like our OT tableaux above, it accounts for the fact that when *Who came?* is answered by *John*, we conclude that *only* John came. However, it also accounts for exhaustive interpretations of explicit partial answers, like disjunctive answers like *John or Bill* or an indefinite answer like *A man*. From the latter answer we can conclude by means of exhaustive interpretation that not all men came, an implicature standardly triggered by the ⟨all, some⟩ scale. The analysis also accounts for the exclusive reading of disjunctive sentences: if (15a) is answered by (15b), the latter is interpreted as (15c) after application of our exhaustivity operator:

(15) a. Did John walk?
b. John walked or Mary walked.
c. John walked or Mary walked, but not both.

Because the relevance relation '>' need not come down to entailment, our exhaustivity operator can account for phenomena Groenendijk and Stokhof cannot account for. First, it has no problems with answers given to *mention-some* readings of *Wh*-questions as discussed in Section 6.4.5. In those cases we predict that exhaustification has no effect. Second, the ordering relation on which we base our analysis of exhaustivity might come down to, for instance, autographic prestige, which means that the examples in (12) can also be handled correctly.

Notice that our exhaustification analysis not only predicts intuitions standardly accounted for in terms of the *Q*-principle; also some *I*-implicatures are accounted for. Just like for Groenendijk and Stokhof's operator, we predict that if the question is *Who quacks?*, the answer *Every duck quacks* is predicted to imply that every quacker is a duck. Horn (2000) calls this inference *conversion* and explicitly proposes to account for it in terms of the *I*-principle. Something similar holds for the inference from *if* to *if and only if*.

Studying Horn (1984) and Levinson (2000) carefully, one sees that two very different kinds of inferences are supposed to be accounted for in terms of the *I*-principle. On the one hand, we have the *strengthening* inferences as discussed directly above, from *if* to *if and only if*, for example. More typical *I*-implicatures, however, are inferences from a sentence to its *stereotypical*, or most probable, interpretation. As we will see, we can capture these *I*-implicatures by means of an operator that is very close to our exhaustivity operator.

The exhaustivity operator given above is defined in terms of an ordering based on utility. As we saw in Section 6.2.1, however, in special cases this utility ordering reduces to one based on informativity. In that case the exhaustivity operator looks as follows:

$$[[exh]] = \lambda T\lambda P.\{w \in W | P(w) \in T(w) \wedge \neg\exists t \in T(w): \\ \inf(\lambda v[P(w) \subseteq P(v)]) > \inf(\lambda v[t \subseteq P(v)])\}$$

Let us now assume that a sentence S gives rise to a set of possible interpretations in any world, that $S(w)$ denotes this set $\{m_1, \ldots, m_n\}$, and that $[[m]]$ denotes the proposition in which m is true. In that case, exhaustivity comes down to the following:

$$[[exh]] = \lambda S\lambda P.\{w \in W | P(w) \in S(w) \wedge \neg\exists m \in S(w): \\ w \in [[m]] \wedge \inf(\lambda v[P(w) \subseteq P(v)]) > \inf(\lambda v[m \subseteq P(v)])\}$$

But what does this formula mean? In particular, how should we interpret question-predicate P in this case? Well, notice that for standard *Wh*-questions we assume that P just denotes a property from worlds to a set of individuals: the extension is the set of all individuals that have property P in that world. For sentences, we can assume something similar. Suppose S is a sentence like *John killed the sheriff*. We might then assume, for instance, that P is a function from worlds to ways in which John killed the sheriff in those worlds. Let's assume that for any world w, $P(w)$ denotes a set. Suppose that in w, John killed the sheriff in a stereotypical way, that is, by knife or pistol. In that case $P(w)$ denotes the singleton set consisting of the state description saying that John killed the sheriff in this stereotypical way, and $\lambda v\ [P(w) \subseteq P(v)]$ denotes the proposition corresponding with this state description.

Because $\inf(A) > \inf(B)$ if and only if $P(A) < P(B)$, we see that for these special cases our exhaustivity operator picks out the most likely, or stereotypical, interpretation of S. Compare this last formula with Blutner's (1998) formalization of the I-principle in terms of conditional informativity (assuming that $[[S]]$ denotes the set of worlds in which S is true under any interpretation):

$$\begin{aligned} I\text{-principle} = \lambda S.\{w \in\ & [[m]] \mid m \in S(w) \wedge \neg\exists m' \in S(w): \\ & w \in [[m']] \wedge \inf([[m]]/[[S]]) > \inf([[m']]/[[S]])\} \end{aligned}$$

One can see that they differ at two points: (i) whereas our interpretation rule considers only alternative interpretations of predicate P, Blutner allows the alternative interpretations of a sentence to vary in much more unconstrained ways; (ii) whereas Blutner considers *conditional* informativity of the state descriptions after the semantic meaning of S is learned, we consider the informativity of the state descriptions themselves. If we also assume that Blutner allows only for variations with respect to a particular predicate, and if the probability ratios between the elements of S do not change after you learn that S is the case, that is, if we make the following assumption: $\forall m, m' \in S$: $P(m/S) > P(m'/S)$ iff $P(m) > P(m')$, our exhaustivity principle and Blutner's formalization of the I-principle come down to the same. But this suggests that we have come to the remarkable conclusion that both Q- and I-implicatures can, in principle, be accounted for by the same principle of exhaustive interpretation!

8 Bidirectional OT and Horn's division of labor

In the previous section we have reduced both the Q- and the I-principles to the principle that we interpret sentences exhaustively. We saw that this assumes that *speakers* are relevance optimizers. However, doesn't that mean that as a result we have to give up on Blutner's bidirectional OT? In particular, how could we now account for Horn's (1984) division of pragmatic labor, so elegantly explained in terms of Blutner's OT, and so important to explain why marked expressions typically get non-stereotypical interpretations?

The solution to this problem readily suggests itself: we can still make use of bidirectional OT, but we base the theory not on the Q- (or S-) and I- (or H-) principles, but rather on the principles of *relevance maximization* (the R-principle) and *effort* minimalization (the E-principle). We have seen that many *Q*- (and some *I*-) implicatures can be captured by our assumption of relevance maximization. The inference to stereotypical interpretation can be accounted for by the *I*-principle, which should, I believe, be part of the principle to minimize effort. The *I*-principle does not mention alternative expressions. To account for markedness phenomena, however, or Horn's division of pragmatic labor, the *E*-principle should take alternative expressions into account as well.

Notice that when we explain interpretation as a balancing act between relevance and effort, our analysis seems very close to Sperber and Wilson's (1986) analysis of natural language in terms of their Theory of Relevance. However, there is an important distinction: whereas Sperber and Wilson seek to maximize relevance from the *hearer's* point of view, we crucially assume that it is the *speaker* who wants to maximize her relevance. This conclusion, I take it, is very much in accordance with Zeevat's (2000) criticism of Blutner's original formulation of bidirectional OT. Blutner crucially assumed that the hearer wants to minimize his effort to understand what the speaker meant. Zeevat argues forcefully that this gives too much responsibility to the hearer: he just has to find out what the speaker meant. So it seems that just like Sperber and Wilson, Blutner also overrated the responsibilty of the hearer in the interpretation process: both maximization of relevance *and* minimization of effort are primarily important from the speaker's point of view. But if we minimize the role of the hearer in this way, it seems that the understanding of bidirectional OT as appealed to in the introduction to this chapter – as an interpretation game between speaker and hearer – is not as straightforward as it seemed. Indeed, I believe that we should think of bidirectional OT primarily as a theory that explains why certain linguistic conventions – in particular Horn's division of pragmatic labor and some principles of centering theory – typically emerge, and that these general conventions, in turn, explain why participants of a particular conversation say and interpret sentences in the way they do.[26] However, although bidirectional OT should be thought of primarily as a theory of language organization, these organizational principles can only be explained in terms of economical language use.

Acknowledgments. The research of this work is supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (KNAW), which is gratefully acknowledged. I would like to thank David Beaver, Reinhard Blutner, Robyn Carston, Gerhard Jäger, Marie Nilsenova, Katrin Schulz, Henk Zeevat, and the anonymous reviewers for their comments and suggestions. Some ideas developed in Section 6.4 are closely related with some proposals made by Reinhard Blutner (p.c.) and Katrin Schulz (2001). Many thanks also to Raj Singh for commenting on an earlier version of this chapter and for correcting my English.

Notes

1. According to optimality theory there also exists a generation function, G, that assigns to each form f a *set* of interpretations that it could possible mean. For ease of exposition I will ignore this function, but all form–meaning pair combinations that play a role in the definitions will obey this constraint: for all $\langle f, m\rangle$ mentioned, $m \in G(f)$.
2. Dekker and van Rooy (2000) have shown that this notion of optimality can be thought of as a special case of the notion of optimality used in Game Theory: it corresponds with the standard solution concept of a *Nash equilibrium* (in updated games). Also Parikh's (2000) game-theoretical analysis of successful communication is formally very close to Blutner's bidirectional OT. For discussion, see van Rooy (to appear).
3. More in detail, $\inf(m/[\![f]\!])$ is $-log_2P(m/[\![\mathrm{f}]\!])$, where P is a probability function, and the probability of C conditional on B, $P(C/B)$, is ~determined as $\frac{P(B \wedge C)}{P(B)}$. The logarithm with base 2 of n is simply the power to which 2 must be raised to get n. Thus, if $P(C/B) = 1/4$, then $-log_2P(C/B) = 2$, because $2^2 = 4$, and if $P(C/B) = 1/8$, then $-log_2P(C/B) = 3$, because $2^3 = 8$. Thus, in case $P(C/B)$ gets lower, the value of $\inf(C/B)$ gets higher.
4. In fact, Blutner (1998) argues that this reduction is in line with Atlas and Levinson's (1981) and Horn's (1984) reduction of Gricean pragmatics to the two contrary *Q*- and *I*-principles. Katrin Schulz convinced me that Blutner was wrong here. I will come back to this.
5. Beaver's analysis of centering in OT extends the empirical coverage of the theory considerably. I will limit myself to original centering, however, and Beaver's reformulation of it.
6. For simplicity, I will just assume the descriptive adequacy of centering theory, although I am aware that since the original statement of centering theory many alternatives have been proposed.
7. Ignoring the more specific gender/number constraints.
8. For a discussion of some other problems, see Zeevat (2000) and van Rooy (to appear).
9. This analysis of assertions can be extended to questions. See van Rooy (1999, 2002) for details.
10. A collection Q of subsets of W is a partition of W iff (i) the partition covers W: $\cup \mathrm{Q} = W$, and (ii) the elements of Q do not overlap: $\forall q, q' \in Q : q \cap q' = \emptyset$.
11. In fact, we do not have to limit ourselves to partitions, but I will do so to simplify matters.
12. Thus, for all $q \in Q$ it holds that $P(q/B) = \dfrac{1}{card(\{q \in \mathrm{Q} | \, B \cap q \neq \emptyset\})}$.
13. Note that by quantification over probability functions, our ordering relation '>' induced by entropy does not generate a total ordering anymore.
14. Although argumentative value is defined rather differently from entropy value, $EV_Q(\cdot)$, observe that in case of binary issues (is h true or $\neg h$?), the two notions of *irrelevance* coincide.
15. This definition is not exactly the same as the one used by Merin (1999); he in fact uses Good's (1950) function that measures the *weight of evidence*, a function that is continuously monotone increasing with respect to $AV_h(\cdot)$.

16. Compare this also with the strongest meaning hypothesis of Dalrymple et al. (1998).
17. The reason is that, in the end, the presumption of optimal relevance is not stated in terms of optimization of the extent of Condition 1. It is only demanded that this extent has to be 'sufficiently' high. No independent measure of what counts as being sufficient is given, however. If 'sufficiently high' means 'having a positive utility', almost the entire notion of relevance comes down to minimizing processing effort.
18. According to one reviewer, this analysis justifies something weaker than Stalnaker was claiming.
19. Proponents of Sperber and Wilson Relevance Theory won't find this very surprising: Sperber and Wilson (1986) themselves explain such phenomena by appealing to the notion of 'processing effort' which my notion of utility by itself doesn't capture.
20. See Gärdenfors (1988) for an analysis of revision of probability functions.
21. Levinson's (2000) *I*-principle is formulated as follows: "Say as little as necessary; that is, produce the minimal linguistic information sufficient to achieve your communicational ends." According to Levinson (2000) this principle means the following from the hearer's point of view: "Amplify the informational content of the speaker's utterance, by finding the most *specific* interpretation up to what you judge to be the speaker's m-intended point, unless the speaker has broken the maxim of Minimization by using a marked or prolix expression." This suggests taking the maximally informative interpretation, and indeed, he explicitly defines *p to be more specific than q* if (a) *p* is more informative than *q*; and (b) *p* is isomorphic with *q*. Strangely enough, however, the *I*-principle is also supposed to account for the inference to stereotypical interpretations, which by definition are not the most informative at all. It is unclear to me how that is supposed to follow on Levinson's reading of 'specificity'. In this section I will assume that the *I*-principle simply demands selection of the most informative interpretation.
22. Or perhaps just the notion of utility, because it seems reasonable to assume that the second condition of our notion of relevance is already captured by Blutner's notion of *effort* in bidirectional OT.
23. Schulz (2001) proposed this alternative way of interpreting Grice.
24. The result of this tableau can also be captured by the following exhaustivity operator that takes a number and a predicate as arguments and results in a proposition:

    $$Exh(t)(P) = \{w \in P(t) | \neg\exists t' \in P(w) : P(t') > P(t)\}$$

    Note that this exhaustivity operator says that one should interpret the sentence as relevantly as possible. In fact, Zeevat (1994) proposed something like this exhaustivity operator, but with '>' replaced by '⊨'. Thus, according to Zeevat one should interpret a sentence *as informative* as possible.
25. Although problematic, neither Gazdar (1979) nor Soames (1982) actually make this wrong prediction. Gazdar does not make it due to his assumption that the scalar implicatures are not allowed to be inconsistent with the clausal implicatures, and Soames by not weakening the force of scalar implicatures.
26. See my 'Signalling games select Horn strategies' (to appear) for more on this.

9

Remarks on the Architecture of Optimality Theoretic Syntax Grammars

Ralf Vogel

This chapter argues for a particular architecture of Optimality Theory (OT) syntax. This architecture has three core features: (i) it is bidirectional, the usual production-oriented optimization (called 'first optimization' here) is accompanied by a second step that checks the recoverability of an underlying form; (ii) this underlying form already contains a full-fledged syntactic specification; (iii) the procedure checking for recoverability especially makes crucial use of semantic and pragmatic factors.

The first section motivates the basic architecture. The second section shows, with two examples, how contextual factors are integrated. The third section examines its implications for learning theory, and the fourth section concludes with a broader discussion of the advantages and disadvantages of the proposed model.

1 Syntax in optimality theory – a proposal

An OT system maps an input to an output according to a system of hierarchically ordered criteria. Such systems can be developed for the modeling of many different things, not only linguistic processes. A central question for the design of an OT system is the choice of the objects serving as input and output and their representational formats. OT systems that use the same objects for input and output have to be distinguished from those that use different ones.

In much of the work in OT phonology, input and output consist of the same elements. For example, the mapping from the input "[tag]" to the output "/tak/" in German describes the process of final devoicing by using strings of phonological segments in both input and output. In their discussion of syllabification, Prince and Smolensky (1993/2002) use output representations that contain the input representations and enrich them with syllable structure. Thus, the plural form for German "/tag/", "[tag + ə]" is mapped onto "/ta.gə/". Other tasks require syllable structure already in the input. One example is the description of loan word integration. Languages that avoid codas and complex onsets resyllabify loan words with such properties. Kenstowicz and Sohn (1998) show this for the Korean dialect of North Kyungsang, where, for example, the name "kris.to" is turned into "ku.ri.su.to."

In OT syntax, a model that has often been used is that of a mapping from a *semantic* representation in the input to a *syntactic* representation in the output (Grimshaw, 1997). Here input and output are radically different. The input–output mapping has the character of a *translation*.

But just as in the case of loan words shown above, it might also sometimes be useful to have the same types of representations, for example, if one wants to describe the typology of syntactic constructions: If language A lacks a particular construction C that occurs in language B, an OT model could show that C would be mapped onto a different construction D if it was in the input in language A.[1]

One example in case is the typology of free relative constructions as modeled in Vogel (2001 and 2002b):

(1) German free relative and correlative construction:

a. Wer	einmal	lügt,	lügt	auch	zweimal		
who-NOM	once	lies	lies	also	twice		
b. Wer	einmal	lügt,	der	lügt	auch	zweimal	
who-NOM	once	lies	that-one-NOM	lies	also	twice	

Free relative constructions (FR) as in (1a) are marked compared to correlative constructions (CR) as in (1b): languages that have FRs also have CRs, but there are languages with CRs that lack FRs. Also, languages with FRs differ in the contexts which allow for this construction – contexts which allow for free relatives also allow for correlatives, but there are contexts allowing for correlatives that do not allow for free relatives. For example, in German, a FR is out, if it would imply the suppression of oblique case (in the following example, dative):

(2)

Wer	einmal	lügt, *(dem)	glaubt	man	nicht
who-NOM	once	lies the-one-DAT	believes	one	not

The solution I proposed in the works cited above is an OT system where the syntactic structure (FR or CR) is specified in the input, and where FRs and CRs compete in the output. In cases like (2), a FR in the input is neutralized to a CR in the output. A CR in the input, however, is always mapped onto a CR in the output.

Another source of the plurality of architectures is the fact that OT syntacticians come from different frameworks. OT syntax work has been done within Government and Binding Theory, Minimalism, Lexical Functional Grammar, Functional Grammar and possibly even more frameworks (representative examples can be found in the collections by Legendre, Grimshaw and Vikner, 2001; Dekkers, van der Leeuw and van de Weijer, 2000; Sells, 2001). These frameworks essentially differ in the character, number and formats of representations that they use.

My impression of current OT syntax work is, nevertheless, that OT systems developed within the different frameworks can usually be translated in a straightforward way without any damage to the systems themselves. The explanatory value of an OT model is usually independent of the representational 'language' that is used. Very often, OT constraints are defined in a quite informal way. This makes the translation from one framework into the other quite easy. In fact, the choice of framework seems to become a minor issue.

This is an expected outcome insofar as the explanatory burden is shifted from assumed properties of representations to constraint interaction. The question of what is the appropriate representation for a particular syntactic construction has less 'weight' within the theory. But this also means that representations can be simplified if one uses OT in explaining syntactic phenomena.

On the other hand, as long as OT syntax work looks so diverse, and is not formulated independently of non-OT frameworks, OT in syntax looks more like a *method* adapted *within* different 'traditional' frameworks than like a framework in its own right. What might be achievable in approaching the latter aim is the development of a kind of 'meta-language' for syntactic representations.

Which representations does an OT syntax system actually need? I want to follow Jackendoff (1997) who summarizes the traditional point of view of what grammars are doing: he claims that there are three representations, a semantic, a syntactic and a phonological representation, and it is their *correspondence* that is modeled by a theory of grammar. Let us use the symbols **M** (for 'meaning'), **S** (syntax) and **P** (phonology) for these representations. The syntactic frameworks mentioned above differ in their assumptions about **S**, its complexity and format, and in how much of **P** and **M** enters the considerations about **S** and its role in grammar.

In Grimshaw (1997), the input can roughly be identified with **M**, it contains argument structural information. Information structural specifications are included in later work of Grimshaw (see Grimshaw and Samek-Lodovici, 1998). The output candidates come close to what is called 'S-structure' in Government-Binding Theory (Chomsky, 1981). S-structure covers some aspects of **P**, namely, morphology and linear order. But prosodic and metrical structure are not represented at all.

Pesetsky (1997, 1998) models a particular aspect of minimalist grammars (Chomsky, 1995) in an OT fashion, namely, the mapping from LF ('Logical Form', an abstract syntactic representation) to PF ('Phonetic Form'), which can be rephrased as the correspondence between **S** and **P**. The empirical coverage of Pesetsky's work is rather small, touching only on aspects of the overt realization of lexical elements, but the model has more general implications. Truckenbrodt (1999) models the correspondence of syntactic and prosodic phrases, Büring (2001), Samek-Lodovici (2002) and Schmid and Vogel (submitted MS) use similar systems in their discussion of the relation between focus and word order.[2] While Pesetsky's OT model is a 'partial' grammar in the sense that it models a mapping from **S** to **P**, without using **M**, the mentioned works on focus use at least the information structural aspects of **M**. The approaches differ in whether **S** is part of the input (e.g., Pesetsky), or part of the output (e.g., Grimshaw). If **M** is the only input representation, then the output is a pair [**S**,**P**], but if **P** is the one and only output representation, then the input must be a pair [**M**,**S**].

These considerations illustrate a common assumption about the role of syntax as *mediating* between 'meaning' and 'sound'. One way of modeling this could be a serialization of two optimizations, one where **M** is mapped onto **S**, and a second step, where the winning **S** is mapped onto **P**. This would imply that there is no direct correspondence relation between **M** and **P**. But the works on focus mentioned above make crucial use of constraints reflecting the correspondence of **M** and **P** – it is uncontroversial that prosodic structure directly reflects information structure. The picture that we get looks more like a triangle: **M** is connected with both **S** and **P**, as are **S** and **P** connected with **M**.

In my work on free relative constructions discussed above (Vogel, 2001 and 2002b) I show the need for having **S** in *both* input and output. The mediating function of **S** is reflected by this double occurrence. The main motivation for this structure, however, is the need to implement a basis for optionality and ineffability of syntactic constructions. In the case of FRs and CRs introduced above, it is obvious that the two constructions stand in a markedness relation: FRs are more marked than CRs, and CRs can always be inserted for FRs, but not always vice versa. The two constructions only differ formally. Universally, the set of languages that have FRs is a proper subset of languages with CRs, and within a particular language, the set of contexts that allow for FRs is a subset of those that allow for CRs.[3]

For a marked structure to survive the competition against the unmarked one, it must be given some advantage, which is usually done by specifying it in the input. Faithfulness constraints ensure that the marked structure wins, as long as the markedness constraints that this structure violates are ranked lower than the faithfulness constraints that are violated by a less marked candidate.

The model that I propose for OT syntax combines two basic issues: having a way of accounting for optionality and ineffability in a standard OT fashion, and implementing the mediating function of syntax. In sum, the structure of input and output (candidates) is the following (the two occurrences of **S** are distinguished by subscripts):

(3) Input and output representations in OT syntax, (see Vogel 2002b):

Input: $\mathbf{S}_I$, **M**
Output: $\mathbf{S}_O$, **P**

The models discussed thus far share the property of being *unidirectional* models. Recent work has suggested that for some purposes a *bidirectional* perspective is necessary. Especially Wilson (2001), Kuhn (2001) and Lee (2001a, 2001b) have to be mentioned here. 'Bidirectional' means here that besides an optimization from meaning to form, OT syntax needs a second optimization from form to meaning. Applications of this idea are still quite rare. Wilson (2001) uses a serial model where optimization from meaning to form restricts the candidate set for the second, syntactic optimization. The model that I argue for in this chapter uses interpretive optimization as a 'post-filter' mechanism. This idea also has predecessors.

Pesetsky (1997, 1998) introduced a constraint that he called Recoverability, which requires semantically relevant material in **S** to be 'visible' at **P**.[4]

But recoverability can only be checked in a process that reverses the direction of optimization: the original output serves as input, and the original input should be the optimal output of the former's optimization. If this is the case, then recoverability is proven. Lee (2001a, 2001b; see also Beaver and Lee, Chapter 6) shows that not only semantic aspects are subject to the recoverability condition, but also syntactic ones. An underlying object-subject order might not be recoverable from **P**, if subject-object order is the unmarked case in a language, and if there are no morphological or other hints that signal the underlying marked order – a classical case of neutralization. The following German example combines the two aspects of recoverability:

(4)	Zwei	Professoren	haben	drei	Studenten
	two	professors	have	three	students

The default interpretation for a clause like (4) is that it has subject-object order and a quantifier scope that follows the linear order of the quantifiers. However, the two NPs are ambiguous for nominative and accusative, and object-subject order is not ungrammatical in principle. Likewise, scope reversal would be possible under other circumstances, or with the help of contextual factors.[5] Thus an input that is specified for object-subject order and inverse scope relations should be able to survive. That the structure in (4) does not have this interpretation in the default case results from a second step of optimization. In this second step, we are looking for the optimal underlying structure of a given surface form. Here the input is the winning **P** of the initial optimization process and we look for the optimal underlying pair [**S**, **M**]. I call this second step *feedback optimization* (see Vogel, 2002). This grammar has the following structure:

(5) Input and output representations in bidirectional OT syntax:

First optimization:	Input: $\mathbf{S}_I$, **M**
	Output: $\mathbf{S}_O$, **P**
Feedback optimization:	Input: **P**
	Output: $\mathbf{S}_I$, **M**

The model emphasizes the role of **P** as the *ultimate* representation in terms of which all underlying information, both semantic and syntactic, has to be encoded. **P** includes *all* aspects of the 'surface form', in particular, it is also the only representation that encodes linear order. This is a common assumption in contemporary generative syntax (cf., e.g., the work based on Kayne, 1994). In these models, the abstract syntactic representation only encodes dominance and relations derived from this, like constituency and c-command, furthermore, it contains the abstract features of lexical items, and syntactic categories.

(6) Assumed representations and what they represent:

M: argument structure, scope relations, information structure etc.
S: constituency, abstract features, syntactic categories etc.
P: linear order, overt morphology, prosodic structure etc.

There are many 'natural' ways of *encoding* relations within these representations. For example, the semantic relations quantifier scope and argument structure are usually translated into (asymmetric) c-command at **S** and precedence at **P**. Likewise, predication is encoded into sisterhood at **S** and adjacency at **P**. Assuming corresponence constraints that formulate these 'default translations' is straightforward. We will turn to some examples in the next section.

It is crucial that the same constraint hierarchy is used in both optimization steps. The recoverability condition is implemented into this model as a condition on grammaticality:

(7) Grammaticality:

A triple [$\mathbf{M}_i$,$\mathbf{S}_i$,$\mathbf{P}_i$] is grammatical, if and only if the input [$\mathbf{M}_i$,$\mathbf{S}_i$] yields [$\mathbf{S}_i$,$\mathbf{P}_i$] in first optimization, and the input [$\mathbf{P}_i$] yields [$\mathbf{M}_i$,$\mathbf{S}_i$] in feedback optimization.

Ungrammaticality may arise in both optimization steps. An $\mathbf{S}_I$ might be mapped onto a different $\mathbf{S}_O$ in the first optimization – ungrammaticality of a particular syntactic structure; or $\mathbf{S}_I$ wins the first optimization, but loses the feedback optimization of its winning **P** – a case of unrecoverability under particular circumstances, usually connected to indeterminacies given in the surface form. The next section discusses example applications of this model. It will also show that the model may not be viewed as 'encapsulated'. Especially, markedness constraints on **M** have to make crucial use of information provided by context and world knowledge.

2 Two examples

2.1 Word order freezing

Let us first consider a simple case of word order freezing in German:

(8) a. Den Hans liebt Maria
the-ACC H. loves M.
'As for Hans, Maria loves him'
b. Hans liebt Maria
H. loves M.
'Hans loves Maria'

Both "Hans" and "Maria" are ambiguous for nominative and accusative case in (8b). Without contextual disambiguation (8b) cannot be interpreted like (8a). The unmarked case is subject-object order. A marked order requires disambiguation, in (8a) the determiner marks the initial NP as accusative. The fronting of "den Hans" reflects the topic status of that NP.

I will now reconstruct this case using the following constraints on the *correspondence* of **M** and **S**:

(9) Constraints on M⇨S mapping:

(elements of **M** are called, 'm_n', elements of **S**, 's_n', and elements of **P**, 'p_n'; identical indices indicate correspondence of elements, e.g., m_1 corresponds to s_1)

a. ARG⇨**S**: If an argument m_1 is higher than another argument m_2 at **M**, then s_1 asymmetrically c-commands s_2 at **S**.
b. INF⇨**S**: If m_1 is [+topic] and m_2 is [−topic] at **M**, then s_1 asymmetrically c-commands s_2 at **S**.

These two constraints conflict in the case of (8a), where the lower argument, the object, is topic. That this clause is grammatical, shows that the order of the two constraints in German must be:

(10) INF⇨**S** >> ARG⇨**S**

If the ranking was the other way around, then such a structure could not survive the first optimization: it would lose against a subject-initial structure. In feedback optimization, we have **P** in the input and search for the optimal underlying form, a pair [**M**,**S**]. Here, the only difference between (8a) and

(8b) is important: the determiner, which signals the case of the initial NP. The correct 'translation' of the surface morphology into underlying abstract syntactic features is evaluated by a constraint on **S⇨P** correspondence. The bare noun "Hans" fits both nominative and accusative, so neither of these two 'interpretations' would violate **S⇨P** for (8b). Likewise, the initial NP "Hans" can be interpreted as topic, independent of its grammatical function, INF⇨**S** cannot be decisive either, and so finally ARG⇨**S** makes the decision favouring a subject-initial structure:

(11) Feedback optimization for (8b):

Hans liebt Maria	S⇨P	INF⇨S	ARG⇨S
OVS, O=topic			*!
☞ SVO, S=topic			

But in the case of "den Hans" in (8a), **S⇨P** is violated by the candidate that interprets this NP as nominative instead of accusative, and so the OVS candidate is the winner:

(12) Feedback optimization for (8a):

Den Hans liebt Maria	S⇨P	INF⇨S	ARG⇨S
☞ OVS, O=topic			*
SVO, S=topic	*!		

S⇨P is an interesting constraint, because its classification as faithfulness or markedness constraint is different in the two optimization steps. Markedness constraints only evaluate properties of candidates irrespective of the input. In this respect, **S⇨P** behaves like a markedness constraint in the first optimization.[6] In feedback optimization, **P** is in the input and **S** in the output. **S⇨P** now acts as a faithfulness constraint.

Another important aspect of this perspective on grammaticality is its context dependency. The effects of word order freezing can be overcome. In the context of a question like (13), the preference for the interpretation of (8b) is clearly object-subject order.

(13) Wen liebt Maria
who-ACC loves M.

Let us assume that the context, a discourse representation of whatever format one prefers, is present and accessible for constraint evaluation. We can then formulate a constraint like (14):

(14) **M**fits**C**: **M** is compatible with the context **C**

This constraint is a markedness constraint on possible interpretations. It favors interpretations that fit into a given context over others that do not fit. It only plays a role in feedback optimization, as only here **M** is part of the candidates and therefore subject to evaluation. The constraint plays the same role as **S⇨P** in the example we had before, in preserving the marked underlying OVS order:

(15) Feedback optimization for (8b) in the context (13):

Hans liebt Maria	MfitsC	INF⇨S	ARG⇨S
☞ OVS, O=topic			*
SVO, S=topic	*!		

Likewise, such a preference can be triggered by world knowledge, as in (16), where only the second NP can meaningfully be interpreted as having the experiencer role of *love*:

(16) Fussball liebt Maria
football-NOM/ACC loves M.-NOM/ACC
'Football, Maria loves'

Let us assume that another markedness constraint on **M** plays the decisive role here, which is similar to **M**fits**C**. It can roughly be formulated as '**M** fits the world'.

This model of grammaticality assumes that we use all resources we can in order to recover underlying structure. At least the second step of optimization is non-encapsulated, and in this respect the model differs from the traditional generative grammarian point of view.

This is not a model of semantic interpretation, it is a model of grammaticality. But it makes use of semantic and pragmatic factors, because it assumes that these factors are crucial for grammaticality to a certain extent.

Grammars may differ in the role pragmatics plays for grammaticality. For Russian, which allows for object-subject orders in principle, it has been claimed that a clause like (16) is ungrammatical under case ambiguity in "non-emotive speech" (see Bloom, 1999). World knowledge obviously does not help in escaping word order freezing in Russian, which would mean that the respective constraint is ranked lower than ARG⇨**S**.

2.2 Superiority and discourse-linking

The paradigm in (17) displays a well-studied contrast in the syntax of English multiple questions:

(17) a. *What did who do?
 b. What did which student do?

This contrast has been discussed in detail by Pesetsky (1987). His explanation for the difference between (17a) and (17b) is that (17b) is grammatical, because the *which* NP is what he called 'discourse-linked' (d-linked): it refers to a set of individuals that has already been introduced in the preceding discourse. This, we infer from this argument, does not hold of *who* in (17a). But Bolinger (1978) has already shown that the empirical generalization about (17a) is also not as straightforward as people often think. He gives the example in (18) to show that this clause can be acceptable in a suitable context (capital letters indicate main stress):

(18) I know what just about everybody was ASKED to do, but what did who (actually) DO?

This example strengthens Pesetsky's point: here, *who* refers to individuals that have already been introduced into the discourse, and the clause is acceptable. The scenario that I want to reconstruct in this section has the following features:

- There are two forms, *who* and *which*:
  - *who* is interpreted as non d-linked by default, but can be interpreted as d-linked *given* the right context;
  - *which* is interpreted as d-linked.
- Both elements are individual lexical items and as such can be part of the input.
- The two elements are related on a markedness scale: *which* is more marked than *who*.

This case is an example of 'partial blocking': *who* could be interpreted as d-linked, but the mere existence of *which* usually blocks it. Under particular conditions, however, this blocking can be overcome. *Who* is assumed to be the unmarked form, because it goes along with non d-linking, which seems to be the unmarked interpretation, though it is not the only one possible. *Which* can only be interpreted as d-linked. So we have two markedness scales:[7]

(19) a. *who, what…* < *which* NP
b. −d-linked < +d-linked

These two scales can be used for the generation of constraints with the method of 'harmonic alignment', developed by Prince and Smolensky (1993/2002). In a first step, we build two subhierarchies of constraints, one for each form ('dl' is an abbreviation for 'd-linked'):

(20) a. **who*/+dl >> **who*/−dl
b. **which*/−dl >> **which*/+dl

The two rankings in (20) are universally fixed, but their interaction is free.[8] Suppose that the ranking in English is the following:

(21) **which*/−dl >> **who*/+dl >> **who*/−dl >> **which*/+dl

(21) states, for instance, that the most marked case is the one where *which* is interpreted as non d-linked. This is the only case that is not attested in English, as far as I can see. I assume that, although *who* and *which* are already specified in $\mathbf{S}_I$, they nevertheless compete in candidate sets.

The non-occurrence of non d-linked *which* can be prohibited with a constraint on input preservation in **S**, $\mathbf{S}_I$⇨$\mathbf{S}_O$. It is ranked below **which*/−dl:

(22) **which*/−dl >> S_I⇨S_O >> **who*/+dl >> **who*/−dl >> **which*/+dl

The predictions of this system are easy to detect: a [−dl] *which* input yields *who* as output. In all other cases, the output form is the one given in the input:

(23) First optimization:

Input: *which*, +dl → *which*
Input: *which*, −dl → *who*
Input: *who*, +dl → *who*
Input: *who*, −dl → *who*

In feedback optimization, we take the form we obtained as input and look for the best interpretation, that is, either d-linked or non d-linked. As there is no faithfulness involved here, it is clear that *who* yields [−dl], and *which* yields [+dl]:

(24) Feedback optimization:

who → −dl
which → +dl

Our model of grammaticality combines the two perspectives, and treats as grammatical only those [input, output] pairs where the input is recoverable from the output. Only two of the four cases in (23) have this property, namely, (25a) and (25d):

(25) First plus feedback optimization:

a. Input: *which*, +dl → *which* → +dl
b. Input: *which*, −dl → *who* → −dl
c. Input: *who*, +dl → *who* → −dl
d. Input: *who*, −dl → *who* → −dl

This system derives the default interpretations that we observed for the *Wh*-phrases under examination. One reading is missing, namely, the contextually forced [+dl] interpretation for *who*, as exemplified in (18). It will be preserved if contextual information is taken into account. To include this, we introduced the general constraint 'MfitsC' in the previous section which may also be used here. It is ranked on a par with $\mathbf{S}_I$ ⇨ $\mathbf{S}_O$:

(26) **which*/−dl >> MfitsC $\mathbf{S}_I$⇨$\mathbf{S}_O$ **who*/+dl >> **who*/−dl >> **which*/+dl

Feedback optimization within the right context gives *who* the chance to be interpreted as [+dl] (27a):

(27) Feedback optimization, including context:

a. *who*, context: +dl → +dl
b. *who*, context: −dl → −dl
c. *which*, context: +dl → +dl
d. *which*, context: −dl → +dl

The discussion in this subsection demonstrates that harmonic alignment can implement the 'division of pragmatic labor' (Horn, 1984), the observation that unmarked forms tend to be used for unmarked situations and marked forms for marked situations. Harmonic alignment can be an alternative to 'weak bidirectional systems' (see also Beaver and Lee, Chapter 6) in the sense of Blutner (2000). The most important effect of a weak bidirectional system – modeling of the division of pragmatic labor – can be implemented within a strong bidirectional system like the one developed in this article. One prerequisite for this possibility is that the forms and interpretations in question can sensefully be compared in terms of a single parameter of markedness. For the standard example discussed in Blutner (2000), this is the case. The example is:

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

The two clauses differ in meaning: (28b) has an interpretation where the causation is much more indirect than in the case of (28a). The two markedness scales that we can use for harmonic alignment here are:

(29) a. $[_{VP}\ V] < [_{VP}\ V\ [_{VP}\ V\ ]]$
(‘simple VP is less marked than complex VP’)
b. direct causation < indirect causation

Using these scales, we can construct constraints as exemplified above, get a fixed ranking in the desired way and derive the wanted effect.

3 Bidirectional OT syntax and learning theory

The bidirectional model of OT syntax that has been developed in the previous sections is reminiscent of models that have been explored in OT learning theory. Tesar and Smolensky (2000) describe the learning of an OT system as the iterated application of a three-step process in the following way:

(30) The Constraint Demotion/Robust Interpretive Parsing (CD/RIP) OT learning procedure (after Tesar and Smolensky, 2000, p. 62):

Given an overt form **OF** and an (initially arbitrary) constraint ranking, **H**:

a. The learner assigns to **OF** a structural description SD_I including an underlying form **UF**.
b. The learner then applies production directed optimization to **UF** and yields another structural description $\mathbf{SD}_P$.
c. If $\mathbf{SD}_P$ is identical to $\mathbf{SD}_I$, then *H* does not need adjustment.
d. If $\mathbf{SD}_I$ and $\mathbf{SD}_P$ differ, then an error has occurred, the learner needs to adjust **H**. She assumes $\mathbf{SD}_I$ to be correct and applies.
e. Constraint demotion, with $\mathbf{SD}_I$ as winner and $\mathbf{SD}_P$ as loser: constraints that are violated (more often) by $\mathbf{SD}_I$ are reranked below constraints that are violated by $\mathbf{SD}_P$.

It needs to be shown that the OT syntax model proposed here fits into this general description of a learnable OT grammar. What I called ‘feedback optimization’ can be identified as the initial step (30a) in Tesar and Smolensky’s (2000) learning procedure. **P** would then be the overt form, the current constraint ranking would be used to get an interpretation for that overt form, a pair [**S**,**M**]. However, the overt form in that model is a ‘surface reflection’, only the overt part of the winning candidate, and as such, it cannot be subject to constraint evaluation, unlike **P**.[9] Thus, the overt form cannot be **P** itself, but only its ‘reflection’. **P** is part of the structural description of a clause, as well as **S** is.

The interpretation $\mathbf{SD}_I$ should then be identified with the triple [**M**,**S**,**P**]. It contains the underlying form **UF** = [**M**,**S**]. The second step in the algorithm applies production oriented optimization, my 'first optimization', to [**M**,**S**], yielding a structural description $\mathbf{SD}_P$ = [**S**,**P**]. Step (30c) needs slight revision. $\mathbf{SD}_P$ cannot be identical to $\mathbf{SD}_I$, because the latter is a triple [**M**,**S**,**P**], while the former is a pair [**S**,**P**]. Hence, instead of the identity of $\mathbf{SD}_P$ and $\mathbf{SD}_I$, we have to check for the identity of the relevant parts of the two representations. This is in fact the only adjustment that would have to be made, and it appears rather harmless to me. Of course, the major underlying assumption of the whole approach is that the representations we are dealing with are quite complex objects. But this is fairly uncontroversial in the area of syntax.

Tesar and Smolensky (2000, p. 63) mention three scenarios where the algorithm fails. These are the following:

- **Selecting an interpretation that cannot possibly be optimal**. This can happen with 'weird' optimal forms which are highly marked. The learner nevertheless assigns an interpretation to it. But this interpretation will not survive the second optimization process. This causes reranking, which then causes, in the next cycle, a new interpretation for the overt form, which again does not survive, again constraint demotion applies and might reestablish the ranking we had before, and the system might run into an endless cycle till it stops.
- **The optimal interpretation is harmonically bound**. A winning interpretation is found to lose under any ranking in the second step of optimization. This situation is easy to handle: the learner can give up learning on the particular data. There is no ranking that would derive the current interpretation as winner. The grammar cannot be learned with the particular data at hand.
- **Endless alternation between different overt forms**. This is another kind of endless circle. Two different data require different rankings, and trigger these whenever they are processed.

As Tesar and Smolensky already discussed, these situations are rather special. The second problem should not pose particular difficulties as long as it only rarely occurs within the set of training data. The first and the third problem point to possible inconsistencies in a language or the given data. Especially the third case is one where alternatives to strict ranking are usually considered, like, for instance, constraint ties or parallel grammars. Each of these cases might occur as well in syntax learning. For successful learning, it is important that cases like these are rare among the training data.

One further problem could be the acquisition of underlying forms. It is especially problematic in morphology, that is, in the acquisition of 'irregular' lexical items, which have to be acquired as whole *paradigms*, not as single elements, crucially because of allomorphic variation. However, for OT syntax it has usually been assumed that underlying forms are universal, therefore not needing to be learned. For **M**, this is quite clear. For **S**, this is a debatable assumption among syntacticians. The generative tradition assumes that abstract syntactic structures are universal: this includes the inventory of syntactic categories and features, as well as the mechanisms of their combination into larger units. At least one proposal has been put forward recently, Croft's (2001) 'Radical Construction Grammar', that assumes that syntactic constructions are language particular, and thus have to be learned, just like lexical elements have to be learned. This is something that the model proposed here might also be able to live with, as long as constructions can be shown to be as learnable as lexical items in general. This task is beyond the scope of this chapter, however.

4 Conclusion

Beaver and Lee (Chapter 6) discuss different OT architectures and compare how they are able to deal with a number of phenomena. The model for OT syntax developed here belongs to their category of 'strong bidirectional models'. Beaver and Lee show that models of this category can successfully deal with freezing, blocking, uninterpretability and ineffability, but that they also fail in dealing with optionality, ambiguity and partial blocking. The model that I developed here, interestingly, is more successful in each of these three cases. Section 2.2 showed how at least simple cases of partial blocking can be dealt with by using the method of harmonic alignment. In accounting for the optionality of forms, I formulated the need for a 'double occurrence' of syntactic specifications in both input and output. A marked form specified in the input is preserved in the output by highly ranked faithfulness constraints.

A more difficult case is the ambiguity of a single form. A very hard case that has not been discussed in this chapter yet, is context-independent ambiguity. A potential example is (31):

(31) Welche Frau hat Hans gesehen?
which woman-NOM/ACC has H.-NOM/ACC seen?
'Which woman saw Hans?' OR 'Which woman did Hans see?'

Although German observes freezing with two ambiguous proper nouns, the structural ambiguity is preserved if (only!) one of the two NPs is a *Wh*-phrase. The way out of this problem that I proposed in earlier work (Vogel, 2002a) is redefining the constraints on syntactic ordering such that they only apply to elements of the same syntactic type. Thus, a constraint like 'ARG⇨**S**' would not be violated by any interpretation of (31), because the two NPs are of a different type. One possible way of accounting for ambiguity is thus ensuring that the constraints make no decision between two candidate interpretations, by defining the constraints accordingly.

In Section 2, I showed how partial blocking in the case of word order freezing and simple *Wh*-elements can be overcome by referring to properties of the context. The claim is that contextual factors can uncover the underlying ambiguity of an expression. A well-known example from phonology, which has been discussed by Zeevat (2000) (see also Beaver and Lee, Chapter 6), results from the phenomenon of final devoicing in languages like German and Dutch. In Dutch, the phonetic string [rʌt] is ambiguous for the underlying forms /rʌd/ ('wheel') and /rʌt/ ('rat'). However, in 'real life' the two interpretations can usually be quite easily distinguished by contextual means. Once this context dependency is reflected in a grammar, in the form of constraints like 'MfitsC', there is a way to derive and predict the possibility of two or more interpretations of an expression.

I hope to have shown that such a reflection of pragmatic factors *within* an OT model of syntax is necessary and desirable. Syntax is much less encapsulated and 'autonomous' than generative grammar usually assumes. The discussion in Section 2 suggests that the application of core syntactic constraints is restricted by pragmatic constraints. The picture of grammar that emerges from the considerations in this chapter is that of a 'total grammar' where expressive and interpretive constraints collaborate and interact, and even syntax can only be understood from the perspective of this very global interaction. In turn, a pragmatic principle like the 'division of pragmatic labor' describes the mutual dependency of related meanings and forms. It receives a natural expression within bidirectional OT models.

Acknowledgements. I thank Reinhard Blutner and Henk Zeevat, as well as the participants of the 2002 "Optimality Theory and Pragmatics" workshop for helpful comments, suggestions and discussions. This work has been supported by grant Nr. FOR-375/1-A3 from the Deutsche Forschungsgemeinschaft (DFG) for the research project "Minimality in Optimality Theory" as part of the interdisciplinary research group "Conflicting Rules in Language and Cognition", University of Potsdam.

Notes

1. The first to propose a model with such properties for OT syntax, were Baković and Keer (2001), as far as I know.
2. It is not accidental that much of recent work in OT syntax is devoted to very 'surfacy' aspects of syntax. Radical surface orientation was the major change that OT induced in phonology. Proponents of this surface orientation, in addition to those researchers mentioned in the text, are Geraldine Legendre and Stephen Anderson (see, for example, Legendre, 2001, and references cited there; and Anderson, 2000).
3. This situation is fully parallel to typical cases of markedness in phonology. Consider, for instance, the relation between voiced and voiceless obstruents. All languages that have voiced obstruents also have voiceless obstruents, but there are languages with voiceless obstruents that lack voiced ones. Second, the contexts

where voiced obstruents occur are very often more limited than those for voiceless ones. In German, for example, voiced obstruents only occur in the onset, but never in the coda of the syllable. Voiceless obstruents can occur in both positions. The syntactic example given in the text is only one among many others that could also have been chosen: passive vs. active, object-subject orders against subject-object orders, complementizer-less subordinate clauses vs. complementizer-introduced clauses in English and German, etc.

4. The definitions Pesetsky gives for the RECOVERABILITY constraint, are quite informal:

> A syntactic unit with semantic content must be pronounced unless it has a sufficiently local antecedent.
>
> (Pesetsky, 1998, p. 342)

> This fact is accounted for by a principle called the Recoverability Condition – the idea being that the semantic content of elements that are not pronounced must be recoverable from local context.
>
> (Pesetsky, 1997, p. 154)

5. One possible way of triggering scope inversion would be a question of the following form:

(i)	Wieviele	Studenten	sind	bei	zwei	Professoren?
	How many	students	are	at	two	professors?

6. To be precise, **S⇨P** should be called $\mathbf{S}_O$⇨**P**. The role of $\mathbf{S}_I$ must be restricted to constraints that belong to the $\mathbf{S}_I$⇨$\mathbf{S}_O$ family.
7. The terms *who* and *which* as used in this 'universal' markedness scale should be understood as 'placeholders' for abstract universal functional categories.
8. This means that if we have two fixed subrankings 'A1 ≫ A2' and 'B1 ≫ B2', there are six possible rankings:

(32) a. A1 ≫ A2 ≫ B1 ≫ B2
b. B1 ≫ B2 ≫ A1 ≫ A2
c. A1 ≫ B1 ≫ A2 ≫ B2
d. B1 ≫ A1 ≫ B2 ≫ A2
e. A1 ≫ B1 ≫ B2 ≫ A2
f. B1 ≫ A1 ≫ A2 ≫ B2

9. I thank Reinhard Blutner for making me aware of this problem.

10
Variation in Demonstrative Choice in Swedish

Jennifer Spenader

1 Introduction

This chapter deals with variation in the choice between two demonstrative forms in Swedish and discusses the desirability of modeling pragmatic phenomena such as referential choice, in an Optimality Theory (OT) framework. In particular, the influence of several factors on the choice of referential form are investigated: abstractness and animacy of the referent; and antecedent accessibility, which is operationalized as distance to the antecedent or anchor of the referent in the discourse. The first factors have to do with inherent properties of the referent while the second factor has more to do with discourse structure-particular factors. All these factors are linked to the level of activation of referents, generally acknowledged to underlie choice of referential form. In order to investigate the factors and the relative strengths of their influence, a small elicitation experiment was conducted. The analysis of the produced data showed a significant effect for distance to the antecedent or anchor, with animacy and abstractness of the referent also playing important roles.

The second part of the chapter introduces OT constraints on model-form choices. This constraint set, together with relevant input–output mappings and the relative frequency of forms produced in the experiment were fed into the Gradual Learning Algorithm (GLA) in Praat (Boersma, 1998; Boersma and Weenink, 2000) to produce a Stochastic OT grammar (StOT). The learned grammar together with the constraints proposed were then used to successfully generate output forms which were in the same distribution as those produced in the experiment. These results are discussed in relation to the utility and desirability of using OT for pragmatic problems, focusing in particular on the difficulties involved in determining an intuitive and consistent set of constraints and the role of context in the analysis.

2 Adnominal demonstratives in Swedish

Swedish has at least two different adnominal demonstrative forms,[1] both of which also have a corresponding pronominal demonstrative form. Consider the following examples:

(1) den här hunden, det här huset, de här barnen
‘this here dog’, ‘this here house’, ‘these here children’

(2) denna hund, detta hus, dessa barn
‘this dog’, ‘this house’, ‘these children’

(1) illustrates an adnominal demonstrative in the proximal form which is used with the definite form of the noun. I will call this the compound demonstrative form. There is also a distal form, that is, *den/det/de där*, which won’t be studied here. (2) illustrates what I will call the simple demonstrative form. It is used with the unmarked form of the noun.

There has been very little research that compares the two forms, but a few potential differences are generally acknowledged. First, (1) and (2) are often described as dialectical variants, with (2) being more common in west-coast Swedish dialects whereas (1) is considered to be more frequent in all other dialects.[2] However, both forms are also commonly believed to be interchangeable in most situations in both dialects. Second, corpus studies have shown that the two forms differ in distribution in spoken and written language with the simple form believed to dominate the written language. In Fraurud’s (2000) study of a small corpus of Swedish argumentative prose only 42 adnominal compound forms were found, compared with 303 adnominal simple forms. In Lindström’s (2000) study of compound pronominal and adnominal demonstratives in a small corpus of Gothenburg Conversation (*Samtal i Göteborg*), she found 284 compound tokens. I re-examined the corpus and found 571 simple demonstrative tokens, twice as many, partially disconfirming the idea that the compound forms dominate speech. Note, however, that the great number of simple forms may be due to the dialect of the recorded speakers, many of whom were from the west coast.

In terms of function, Lindström (2000) argues that a major use of the compound form is to remind (*påminnande funktion*), an attested function of demonstratives also termed “recognitional use” by Diessel (2000). These are references to referents that are discourse new, but hearer old, and are clearly presuppositional.

Additionally, compound demonstratives may also be used more deictically, a use that is perhaps encouraged by the place adverbials that help compose it, for example *här* or *där* (*here* and *there*).

Are there differences in the use of the two forms or are they merely stylistic or dialectic variants? The most obvious place to look for differences would be among the many factors generally agreed to play a role in choice of referential expression in studies that look at differences between the choice of a demonstrative form over a definite NP or another pronominal form.

3 Factors known to affect choice of referential form

A multitude of factors have been suggested as an explanation for the choice of one referential form over another, but few are categorical.

Factors can be divided into two groups, those that have to do with inherent characteristics of the referent itself, characteristics that are stable across contexts, and those that are related to the referent's role or level of activation in the current discourse. I present several factors below with information about what earlier research has said about these factors related to demonstratives, focusing on adnominal demonstratives in particular, suggesting also how identification of the factors could be operationalized. Note that I assume an underlying representational level where actual anaphor resolution occurs, in the same way as Discourse Representation Theory (DRT; Kamp and Reyle, 1993) and similar theories.

3.1 Abstractness of the referent

Situations, events, propositions and facts are all abstract objects. Ontologically, they are quite different from concrete objects in that they are not individuated in the same way but are "a matter of convention within our conceptual scheme" (Asher, 1993, p. 258) and are therefore more dependent on the manner in which they are introduced and described. One reason to suspect that abstractness plays a role in the choice of referential form is because demonstrative pronominal forms in particular have often been associated with abstract object anaphoric reference, also called discourse deixis (Webber, 1991; Asher, 1993). Many abstract objects are already implicitly present in the discourse in the form of predicated or clausal information that has not been established as an individuated discourse referent in the discourse representation, and therefore these objects do not have an NP-antecedent in the text. Instead, the referential act itself is considered to instigate a process of reification of this already given information into a discourse referent at the level of representation. Alternatively, already established discourse referents to abstract objects can be re-referred to with an NP anaphor. In this latter case, the ontological status of the referent is still abstract but there is no reification process when reference is made to it, and at the textual level these anaphoric expressions can be said to have NP-antecedents.

In English, abstract objects can be referred to with the pronoun *it*, as well as the demonstrative pronouns *that* or *this*. Looking only at cases where reification was necessary, Webber (1991) has argued that unstressed *this* and *that* are more natural referential choices, but that the information from which an abstract object is reified needs to be in focus, which she operationally defines as the right frontier of a discourse tree structure. Even though this early work focused on demonstrative pronouns, abstractness of the referent may affect the use of demonstrative nominals as well. Additionally, the proposed higher frequency of the simple form in written Swedish may be an effect of the greater frequency of abstract referents in written language, in which cases it would be abstractness rather than language genre that is the underlying cause.

It is not immediately apparent how abstractness and concreteness of the referent can be determined. Using antecedent type, for example NP or non-NP, while easy to code, gives only a rough approximation in that it categorizes anaphoric forms according to the resolution process, that is, reification or not, rather than according to the actual characteristics of the referent, that is, abstract or concrete. Reified referents will tend to be abstract, but not all abstract referents will need to be reified. Intuitive definitions that attempt to identify events, situations, propositions and facts become problematic when applied to real data because there seems to be a range from more to less abstract. Often the more specific an event is, the less it is perceived as abstract, making the perceived degree of abstractness context dependent.

3.2 Animacy of referent

Animacy generally isn't explicitly taken up as a factor in referential choice, but it is often implicitly part of many studies because they concentrate on personal pronouns and their alternatives. Animacy interacts with accessibility[3] in that animate individuals are generally more salient than inanimate individuals, they are generally protagonists and often play central roles in a discourse.

3.3 Level of activation of the referent

Many theories of referential form consider the ease with which referents are identified to be related to the level of activation of the referent in the discourse. Often, forms are ranked according to the level of activation typical of their appropriate use, for example the Givenness Hierarchy of Gundel, Hedberg and Zacharski (1993), and the accessibility hierarchy of Ariel (1991) are two well-known rankings.

In these hierarchies demonstrative pronouns, like other pronouns, are placed high on the scale, while demonstrative noun phrases are placed lower, just above definite NPs. Demonstrative noun phrases are believed to refer to referents with a lower degree of accessibility, because of their greater semantic content which aids in the identification of the referent. Accessibility scales differ in their details. For example, Ariel (1991) classifies full demonstratives as mid-accessibility markers. This is meant to reflect a tendency for them to be used with referents that are less accessible than pronouns, including demonstrative pronouns, but with referents that are more accessible than definite NPs. Ariel argues that full demonstratives differ from definite NPs according to how attenuated they are, and the criteria of attenuation is defined as a measure of how long, or how attention-getting a form is; longer, stressed forms are considered more attenuated than shorter forms, and more marked forms are considered more attenuated than less marked forms. Because demonstratives are rarer, and longer than pronominal forms, demonstratives can be considered to be more marked, thus more attenuated, and therefore are more likely to be able to retrieve entities with a lower level of accessibility. According to this idea, the Swedish compound form is more attenuated than the simple form because it is combined with the longer definite form of the noun. On the other hand, the Givenness Hierarchy assumes a cognitive state termed *activated* for appropriate reference with both pronominal demonstrative forms and *this N* forms, while the lower cognitive state of *familiar* is considered sufficient for *that N* forms. All of these forms are, however, considered to be associated with a level of accessibility higher than definites whose referents need only be *uniquely identifiable*.

Using Centering Theory (CT; Grosz, Joshi and Weinstein, 1995) as a theoretical framework Poesio and Modjeska (2002) studied the use of *this*-NPs in a subset of the GNOME corpus, looking at the texts of museum descriptions and medical information. They tested several different definitions of where the antecedents of *this*-NPs tend to come from, and what characteristics they tend to have. CT is a way to measure local coherence in discourse but has frequently been used as a theoretical framework to describe tendencies in referential choice in many corpus studies. Poesio and Modjeska (2002) conclude that *this-NP*s are used to refer to entities which are *active*, but which were not the backward-looking center of the previous utterance. They define active as entities that are either deictic, given in the previous utterance, or an abstract object or plural referent that could be constructed from information in the previous utterance. The backward-looking center of the previous utterance is the most salient entity from two utterances ago that was realized in the previous utterance. In other terms, the uses of *this-NPs* were generally known and accessible entities, but were entities that were not the focus of attention when the utterance was made, and were additionally not the focus of attention in the previous two utterances.

Maes and Noordman (1995) present a very different view of the function of demonstrative noun phrases based on corpus studies of Dutch. They noticed that in almost all of their corpus examples demonstrative noun phrases could be replaced with definite NPs without leading to difficulties in resolution of the referent, so the traditional treatment of demonstrative NPs as a marked form of the definite NP that aids in anaphor resolution did not seem correct. There was, however, a slight difference in meaning. For this reason, they question the view that demonstrative forms have a special identificational function that differs from definite NPs. Instead, they argue that demonstrative NPs can have a modifying affect on the representation of the underlying discourse referent they refer to. Depending on the context and the lexical-semantic relationship to the antecedent, a demonstrative form can classify, contextualize, or attribute new information to the discourse referent it is resolved to, and they illustrate these three modifying functions with corpus examples. In the current study we could ask to what degree each form seems to have these three different modifying affects.

It seems that demonstrative forms, whether they are full nominals or pronouns, are considered more marked than alternative referential choices, for example, definites or pronouns. This suggests that their referents must have a level of activation or accessibility higher than pronouns or definites (e.g., Ariel, 1991; Gundel, Hedberg and Zacharski, 1993), or that they are used with an altogether different function, as Maes and Noordman (1995) propose. Unfortunately, how Poesio and Modjeska's (2002) results for *this-NPs* fit with the idea of a higher level of activation in demonstratives is not clear. It is also common to consider there to be a difference in the typical level of activation between proximal and distal reference, as well as differences between pronominal and adnominal forms.

The above short presentation should have made clear that referential choice is a multi-factor phenomenon. Identifying what factors lead to the use of a demonstrative rather than an alternative referential form is difficult. Distinguishing between two demonstrative forms is likely to be even more difficult because the difference between them is surely smaller. Any difference will probably surface as the factors identified being associated more with one form than the other to different degrees or in different ways. By looking at these factors when the two forms are elicited in a controlled environment we should be able to discover differences if they exist.

4 Choice of demonstrative: experimental study

4.1 Procedure and experimental design

An elicitation task was conducted. Choice of demonstrative was treated as the dependent variable, where the abstractness and animacy of the referent and the distance to the antecedent or anchor were treated as the independent variables.

4.2 Experimental form

Thirty native-Swedish speakers were presented with seven short, three-line stories and asked to complete each story by providing a subject. All stories were presented twice in a mixed order. Via a drop-down menu, subjects were forced to choose one of the six demonstrative forms (e.g. neuter, common gender and plural for each of the two forms) as a determiner and then to fill in the subject. Each sentence already had a predicate. One presentation of the story had a predicate that strongly suggested an abstract subject, while the other had a predicate that strongly suggested a concrete subject. This was done in order to increase the likelihood of obtaining a balanced number of concrete and abstract subjects in order to make it easier to study the effect of this factor on demonstrative choice. Below is an example of one story with the two possible predicates. Predicate A suggests an abstract subject while predicate B suggests a concrete subject.

Story 1

1. Eleverna vantrivs allt oftare i skolan.
2. Många klasser har alldeles för små klassrum.
3. Lärarna i vår trångbodda skola är väldigt upprörda.
4. A.______har blivit outhärdlig. (abstract predicate)
   B.______ måste byggas ut eller så måste studenterna flyttas. (concrete predicate)

1. Students are often unhappy in school.
2. Many classrooms are far to small.
3. Teachers in our cramped school are very upset.
4. A. ______ *has become unbearable.*
B. ______ *has to be added on to, or students need to be moved.*

The stories were all constructed to have a similar discourse structure and are given in the Appendix at the end of the chapter. The first sentence was a general statement about some issue. The second sentence was a comment on the first sentence. Together the first two sentences could be considered to make up a discourse segment. The third sentence described a more specific example of the general situation introduced in the first sentence. This set-up allows the fourth sentence to easily refer either to the general concept or to be a comment on the more specific case introduced in the third sentence. An example of a possible answer to 4A above is *Denna situation* (the situation) and to 4B *Den här skolan* (this school).

In addition, six more stories with the same structure for the first three sentences were each presented twice. These stories were also part of another experiment, and subjects were forced to choose either a simple or a compound demonstrative determiner and were then asked to complete the entire sentence including the predicate.

The task was web-based and subjects reported that it took them between 25 and 35 minutes to complete the entire task.

4.3 Coding the data

Six factors were originally coded in the data: dialect of the subject, the syntactic form of the antecedent or anchor, for example, NP or non-NP, the abstractness of the referent itself, the animacy of the referent, and the distance to the anchor or antecedent. There was no easy way to operationalize the potential modifying affect of demonstratives identified by Maes and Noordman (1995) so this factor was not examined here.

Abstractness of the referent was determined by the experimenter in consultation with another researcher who was also a native speaker of Swedish. Agreement was reached for all items. Examples of words coded as abstract are *questions (frågorna), choices (valmöjligheter), "crampedness" (trångboddhet), intrigues (intriger), increase (ökning)*. Examples of words coded as concrete include *women (kvinnorna), school (skola), column (spalt), newspaper (tidning)*. Animacy of the referent was coded either as animate, including humans and animals, or inanimate.

As discussed in the background given earlier, the notion of accessibility or availability is problematic. The experimental set-up did not make a CT-analysis appropriate.[4] Instead, a simpler notion of distance was used. Ariel (1991) explicitly and CT implicitly consider distance to the last mention of the referent to be a crucial factor affecting the level of activation of that referent. This is easily coded by noting whether or not an entity that can serve as an antecedent or anchor of the referent appeared one, two or three sentences away (−1,−2,−3) in the story. Consider the following example:

Story 2

1. Skvaller på arbetsplatsen kan ha en viktig funktion.
2. Det skapar närhet mellan de som skvallrar genom att ge en "vi-mot-dem" känsla.
3. På vår arbetsplats finns det många romantiska intriger bland de anställda.
4. **Dessa intriger/Det här skvallret** blir föremål för många spekulationer och diskussioner.

1. *Gossip in the workplace can fill an important function.*
2. *It creates a closeness between those who gossip by giving an "us-against-them" feeling.*
3. *At our workplace there are many romantic intrigues between employees.*
4. *These intrigues/This gossip has become the object of many speculations and discussions.*

Dessa intriger (these intrigues) has an antecedent in the previous sentence, whereas *det här skvallret* (this gossip) is a reference to the gossip about the romantic intrigues mentioned in the previous sentence and can be considered an anchor. Both of these are coded as −1, having an antecedent or anchor in the immediately preceding sentence. Textual NPs were considered to be the source of an antecedent if the form produced by the subject could easily have replaced the textual NP without a great change in meaning. NPs that referred to an entity that was in a clear part-of or hyponym relationship with the subject were considered anchors. Referents implied by verbs or verb phrases or other linguistic units in the text were excluded as anchors. This is not an ideal solution but it is conservative. Textual NPs that are synonyms or strongly imply the referent of the anaphoric expression are generally acknowledged to be anaphorically related, while there is little agreement on what status the implied referents of, for example, verb phrases, should have. These responses, including other cases where there was no antecedent or anchor in the story, were also coded as *no*.

4.4 Results

Eight responses were discarded because the test subjects created a sentence with a demonstrative pronominal subject, either by neglecting to fill in a noun in the subject position or by filling this position with a finite verb. Initial examination showed an effect for dialect with west-coast speakers surprisingly being less likely to use the simple form than in the general distribution ($\chi^2 = 6.15$, $df = 1$, $p \leq 0.05$). For this reason, the four west-coast subjects were excluded from further analysis. This left 357 responses for analysis, 226 simple forms (63 percent) and 131 compound forms (37 percent). Initial tests also showed a significant effect for story ($p = 0.0484$). This is an unfortunate result of the small amount of material used. The exact effect and reason why some stories differed was difficult to determine; however, the story proved later not to be a significant factor in predicting the dependent variable, as will be explained below.

For many stories, subjects choose surprisingly similar or identical lexical subjects. Both the compound and simple forms were chosen though often with a tendency for one form to be strongly preferred. Table 1 below presents the simple percentages found for each item. Frequencies that differ from the general proportion of each form in the data as a whole are possible sources of significant differences.

We can see that given an animate referent, there is a significantly greater chance that it will be used with a simple demonstrative form than with a compound form ($\chi^2 = 4.167$, $df = 1$, $p = 0.0412$). On the other hand, if the referent is inanimate, the percentages are almost the same as the

Table 1 Input–Output pairs based on experimental results

Factor		**% simple**	**% compound**
Animacy	Animate	76.3 (42)	23.7 (13)
	Inanimate	60.8 (183)	39.2 (118)
Abstractness	Abstract	64.6 (148)	35.4 (81)
	Concrete	60.6 (77)	39.4 (50)
Antecedent form	NP	66.7 (120)	33.3 (60)
	non-NP	59.6 (105)	40.4 (71)
Distance to antecedent	no	58.0 (103)	42.0 (74)
	−1	74.0 (103)	26.0 (36)
	−2	61.5 (8)	38.5 (5)
	−3	40.7 (11)	59.3 (16)

percentages for the overall use of the form, so there is no significant preference for either form.

Abstractness of the referent shows percentages very close to the values for the distribution of the forms in the data as whole, and this difference is not significant ($\chi^2 = 1.002$, *df* = 1, p = 0.3167). Neither was syntactic form of the antecedent or anchor significant ($\chi^2 = 2.061$, *df* = 1, p = 0.1508). Distance to antecedent, however, was significant ($\chi^2 = 15.63$, *df* = 3, p = 0.01). Examining the percentages, we can see that the more recently the antecedent or anchor of the referent of the subject has been mentioned, the more likely it is to be referred to with the simple form, while referents referred to in the first sentence were more likely to be referred to with a compound form. When there was no antecedent or anchor given (no), there was a greater chance of the simple form over the compound.

However, the significance of each individual factor is not enough information to determine the contribution each factor makes together with the other factors to the realization of the dependent variable. To determine this, a logistic regression analysis was done. A logistic regression model is one type of Generalized Linear Model, and is the most common statistical method used in studying linguistic variation. This method is able to model how each independent variable contributes to the realization of the dependent variable in the context of the other independent variables, and this is why it is considered an appropriate quantitative method for studying variation affected by multiple factors.[5] The basic procedure considers the relative weights of each factor and the statistical significance of each factor for predicting the value of the dependent variable. An iterative procedure known as step-up analysis compares different models of the data. Step-up first selects the factor with the most predictive power, and then adds each of the remaining independent factors in turn to find the combination of factors (a model) that has the best predictive power for the dependent variable. Models are compared with each other. An excellent explanation of logistic regression is given in Paollilo (2002).

The data was run in the GoldVarb 2.0 (Sankoff and Rand, 1999) program and a logistic model of the relative weight for the use of the simple demonstrative was obtained of 0.632. A step-up analysis was performed[6] on the coded data to test the descriptive predictability of taking different combinations of independent variables into account. The first factor chosen by GoldVarb was distance to antecedent. This is the factor that is most predictive of choice of demonstrative form. The second factor chosen was abstractness of the referent.

This result was unexpected given the statistical significance of animacy and the lack of significance of abstractness when treated as individual factors. But a closer examination of the cross-tabulations between animacy and abstractness (not shown) identified several categories where there were zero values, so-called *knockouts*. The knockouts have two causes. The first cause is structural, the set of abstract objects and the set of animate objects are disjunct, that is, no animate being is abstract and vice versa. The second cause is sparse data; there were too few responses that referred to the second sentence (−2) so there are no examples of animate referents with antecedents at that particular distance. This may be a result of the discourse structure of the stories, which encouraged commenting either on the specific example introduced in the third sentence, or commenting on the general topic, which often meant referring to an antecedent in the first sentence.

These two knockouts were dealt with in the following way. Responses with referents to the second sentence (−2) were removed from the data. The categories for animacy and abstractness were recoded in order to remove the structural problem, resulting in three new categories: animate, inanimate-abstract and inanimate-concrete.

The resulting percentages for the data are given in Table 2 below. These differences are significant ($\chi = 52.11$, $df = 2$, $p = 0.001$). We can also now see what effect abstractness of the referent has, an effect that was not visible when animate objects were also included with inanimate-concrete objects. Inanimate-concrete referents have a higher chance of being referred to with the compound form.

The step-up analysis was rerun using the new categories as factors. Once again, GoldVarb selected distance to antecedent or anchor as the most significant factor in determining demonstrative choice, followed by animacy/abstractness. Taken together these factors make the best predicative model of demonstrative form choice (Log-likelihood −222.598, $p \leq 0.020$), adding information about other factors, including story, worsens the predictive power of the model.

4.5 Discussion of experiment results

For all factors it seems that the simple form tends to mark high accessibility while the compound form marks low accessibility. For example, the simple form is associated with animates. Animates were argued to be inherently more accessible than inanimates. The simple form is also more likely to be used with referents that are in the current discourse focus.

On the other hand, referents with the compound form are more likely to be distant, and therefore less activated. Fraurud's (2000) data also showed

Table 2 Results of recoding animacy and abstractness

Factor	**% simple**	**% compound**
Animate	76.3 (42)	23.7 (13)
Inanimate-abstract	64.6 (148)	35.4 (81)
Inanimate-concrete	48.6 (35)	51.4 (37)

similar effects, 92 percent (n = 163) of simple adnominal demonstrative NPs had NP-antecedents in the same or preceding sentence, versus 77 percent (n = 10) of the compound forms. The use of the compound form with less accessible but given items is also consistent with the results for *this-NPs* in Poesio and Modjeska (2002).

Perhaps the compound form is being used with the same type of recognitional function identified in Lindström's (2000) study, but at a local rather than a global discourse level. The preference for the longer compound form with referents at a greater distance can also be explained by Ariel's (1991) claim that higher degrees of attenuation in the referential form are necessary with referents that have a lower degree of accessibility. The preference for the compound form with concrete referents may also have to do with the locative adverbial that is part of this demonstrative, encouraging its use with things that are in the here and now.

There seems to be an interaction between animacy and activation. Examining the cross-tabulations also revealed that almost all reference to animate entities had antecedents in the third sentence (−1 distance). This means that the strong preference for the simple form with animates is equally accounted for by a preference for the simple form with highly accessible referents.

In summary, in the limited material studied here, the simple form is associated with greater accessibility or activation while the compound form is associated with lower accessibility.

5 Modeling variation in referential choice in stochastic optimality theory

Referential choice amounts to preferring one form over another and OT seems a natural form for modeling these preferences. But because referential choice is the result of competing contextual factors, the role of context in an OT production analysis becomes crucial. Exactly how context should be interpreted in an OT analysis is still an open question, but Kuhn (2001b) has argued that optimization should be relative to a fixed context (to the extent that this is determinable). This view then parallels standard treatments of referential choice where contextual characteristics of elements of the input are assumed to be already given. However, there is no guarantee that all discourse participants interpret the context in the same way (i.e., share the same model).

Most traditional accounts of referential choice base their explanation only on faithfulness to an input with a fixed context, see, for example, the referential hierarchies proposed by Gundel, Hedberg and Zacharski (1993) and Ariel (1991). Zeevat (2002) proposes a number of OT constraints to generate a range of referential expressions, but these are also limited to faithfulness constraints.[7] Zeevat's analysis is based on a number of PARSE constraints that refer to saliency, attention, givenness and uniqueness, and he mentions that he considers demonstratives to be parsing for attention and givenness. These characteristics are assumed to be already identified as present or not in the input. The parse constraints are Max faithfulness constraints that demand that the particular input features be realized on the output. The constraints are also unranked with respect to each other. In his treatment the majority of the work is, however, performed by FaithInt, essentially a Dep constraint, a constraint that penalizes candidates whose form suggests or codes for factors *not* present in the input, leading to an input–output mismatch. The different levels of activation are treated as implicational (as in Gundel, Hindberg and Zacharski, 1993). Forms that are lower on the activation scale than the input requires incur parse violations, while forms that are higher on the scale than the input requires incur FaithInt violations. The tableau in (3) is a slightly modified reproduction of one given in Zeevat (2002, p. 80).

(3) Generation of referential expressions:

referent: book, discourse-given but not salient	Parse Salient	Parse Attention	ParseOld	Parse unique	FaithInt
It					*
This					**
this book					*
this book by Anna					*
☞ the book					
☞ the book by Anna					
a book by Anna			*		

Accounts of referential choice that take markedness into consideration are surprisingly lacking, though a key point in OT methodology is that optimization should be the result of interactions between faithfulness and markedness. Centering Theory discussion of different transition types incorporates some ideas that could be considered to represent markedness in a specific context, and a few ideas appear as constraints in Beaver's (to appear)

OT reworking. But in general, referential expressions have been defined as parsing some underlying factors.

5.1 Constraints on referential choice

Because we are only considering two candidate forms, defining constraints is relatively simple if perhaps unrealistically so. I assume that the simple form is associated with high accessibility and inherent salience, while the compound form is associated with low accessibility and low salience. OT constraints must be declarative, and candidate output forms treated as more or less appropriate should be the result of constraint interaction, and not some sort of pre-evaluation with some sort of weighting of factors (see Kuhn, 2001a, for a discussion). However, referential choice often involves a subjective decision that some referents are more or less salient than others, in the context, depending in part on other potential referents. This can only be treated in OT by identifying qualities or factors as associated with a particular referent in a fixed context, and a substantial part of the analysis is moved to the process of determining the input. This is basically what Zeevat (2002) does. Considering the pilot study data the following constraints seem reasonable:

> **Last S Salient** Referents mentioned in the previous sentence are salient.
>
> **ParseLow** Referents given in the discourse but not mentioned in the previous sentence are not salient, and need an attenuate form.
>
> **ParseAnimate** Animate referents are inherently highly salient.
>
> **ParseConcrete** Concrete referents are inherently non-salient.

Distance to antecedent or anchor was the most significant factor in the results. Last S Salient is a constraint suggested in Beaver (to appear) which is violated when a low-accessibility marker is used with a referent mentioned in the previous sentence. ParseLow captures the tendency to prefer the longer compound form with referents that are less accessible, which are referents mentioned in the first sentence (−3) in the coded data. ParseAnimate incurs violations when a low-accessibility marker is used with an animate referent, referents which are considered inherently highly salient, a referential choice that gives mixed signals to the hearer. ParseConcrete works in a similar way, but penalizes the simple form with concrete referents.

The order of the constraints only leads to effects if the input is contradictory. For example, if the referent is animate, given, but not salient, then the effects of ParseAnimate will conflict with ParseLow. As potential markedness constraints consider the following suggestions:

> ***MarkAnim** Incur one violation mark for every referent explicitly marked by form for animacy, but where animacy can be determined from identification of the referent

***MarkLow** Incur one violation mark for each referent marked as having low accessibility if this is determinable from the previous discourse

***MarkHigh** Incur one violation mark for each highly accessible referent that is marked for this if this is determinable from the previous discourse.

The two tableaux in (4) and (5) show how the constraints are affected by typical input data from the experiment.

(4) Input: animate referent that is highly activated:

Anim + high (−1)	ParseAnim	Last S Salient	ParseLow	ParseConc	*MarkLow	*MarkHigh	*MarkAnim
simple						*	*
compound	*	*					

(5) Input: animate referent that has low activation:

Anim + low (−3)	ParseAnim	Last S Salient	ParseLow	ParseConc	*MarkLow	*MarkHigh	*MarkAnim
simple			*				*
compound	*				*		

The main difference between this system and Zeevat's is that the faithfulness constraints are all MAX constraints that penalize output forms which do not parse input characteristics, but there is no Dep constraint equivalent to Zeevat's FaithInt. Instead, candidate forms that realize input characteristics that can be determined by the context are penalized by markedness constraints. Not all languages seem to make all input distinctions (see Gundel, Hedberg and Zacharski, 1993), and modeling this difference by markedness constraints seems motivated. Note that none of the factors referred to in the constraints implicates the others, another difference from Zeevat (2002).

5.2 The stochastic OT grammar

Stochastic OT (StOT)[8] differs from standard OT in that it uses a continuous scale of rankings rather than a strict ordinal scale as in standard OT, which makes it possible to successfully model variation (e.g., Jäger, Chapter 11; Bresnan, Dingare and Manning, 2001; Lee, 2002). Each constraint is given a value on a scale of real numbers called its *ranking value*. This means that the constraints themselves are not merely ordered with respect to each other, but can also be at varying distances to each other. The actual value of a constraint used in the evaluation of a particular input is a value taken from the normal distribution around the ranking value of the constraint. This value is called the *selection point* of the particular evaluation. This means that constraints with close ranking values may sometimes be evaluated with selection points that reverse the dominance relationship between them, leading to a degree of variation in the output. Constraints ranked far from each other will seldom, if ever, switch rankings and thus will perform as if they were categorical, and a difference of ten units is an almost categorical distinction. The program Praat comes with an implementation of the Gradual Learning Algorithm (GLA) which can learn a stochastic OT grammar when provided with a set of constraints, a set of relevant input–output pairs coded for how they are evaluated with respect to each constraint, and frequency distribution data for each input–output pair.

There are two advantages in using StOT and the GLA. First, it makes it possible to learn a constraint ranking from experimental data that has variation, which can only with great difficulty be done by hand. Second, the learned grammar can be used to generate forms, and the distribution of these generated output forms can be compared with the actual experimental data. If the set of constraints is incomplete, lacking reference to some critical factor, then the proportion of the generated forms will in most cases differ from the input data. This means that the program can be used to evaluate a proposed constraint set.

Six relevant input–output pairs[9] that illustrated the combination of the three relevant features identified in the recoding of the experiment were identified. Examples with accessibility of −2 were excluded. This was because of sparse or non-existent data for some category combinations and because the experimental analysis also showed no difference between the choice of forms for the few examples produced. For each pair, the constraints violated were coded. The actual experimental data was used for the frequency distribution data. (These numbers were given in Tables 1 and 2 in Section 4.4.) Each constraint was given an initial ranking of 100 and the Gradual Learning Algorithm was run twice with 1,000,000 inputs.[10] Learning resulted in the following ranking order:

(6) Ranking order of constraints learned:

LAST S SALIENT	101.203
PARSEANIMATE	100.743
*MARKLOW	100.194
PARSECONCRETE	100.146

ParseLow	99.806
*MarkAnim	99.257
*MarkHigh	98.797

Below are two examples of how the above grammar evaluates two different inputs. Note that for phenomena with the high degree of variation that we see here, it is not illustrative to show several tableaux because the same input will be evaluated with several different constraint rankings:

(7) Input: animate and highly activated referent:

anim + high	Last S Salient	ParseAnim	*MarkLow	ParseConc	ParseLow	*MarkAnim	*MarkHigh
☞ simple						*	*
compound	*!	*					

(8) Input: concrete referent with low activation:

con + low	Last S Salient	ParseAnim	*MarkLow	ParseConc	ParseLow	*MarkAnim	*MarkHigh
☞ simple				*	*		
compound			*!				

The constraints here cluster around the same values and the greatest distinction is only two ranking units. This is typical of forms that exhibit variation and in Boersma and Hayes (2001) most variable forms were the result of constraints ranked between one and two units apart. Because the two forms allow variation for each relevant factor covered by the constraints, we end up with no categorical cases.

The GLA tries to learn a grammar that will generate each of the output forms in approximately the same proportions as the frequency distribution it learns from. Boersma and Hayes (2001) present the generation of correct

Table 3 Actual distribution of output from learned grammar (bold) with the distributions from the experiment data (in parentheses)

Factor		**% simple**	**% compound**
Animacy/ Abstractness	Animate	**72.6** (76.3)	**27.4** (23.7)
	Inanimate-abstract	**67.8** (64.6)	**32.2** (35.4)
	Inanimate-concrete	**46.8** (48.6)	**53.2** (51.4)
Distance to antecedent	−1	**72.4** (74)	**27.6** (24)
	−3	**52.4** (40.7)	**47.6** (59.3)

distributions as evidence that the GLA works. Thus generated forms can be compared with the input and can be seen as a method for evaluating the ability of a constraint set to account for the data.[11] This is especially useful when working with pragmatic constraints, where what makes a good constraint is not at all as clear as it is in, for example, phonology. In Table 3 we can see the frequency of each output form given the factors in the first column.

For all but one factor combination, the grammar produces forms in proportions very similar to the experimental data. The only discrepancy is that the compound forms for referents that have an antecedent with low accessibility (−3) should be preferred to a greater degree than the generated grammar shows. However, the difference between the generated forms and the experiment data is *not* significant ($\chi^2 = 3.021$, $df = 1$, $p < 0.10$).

6 Discussion

In conclusion, given the data studied here there seem to be some subtle, yet statistically significant differences between the two demonstrative forms and it is possible to model these differences in stochastic optimality theory. The simple form seems to be used with more accessible and salient referents, while the compound form shows characteristics of being a stronger form, appropriate for referents, with a lower level of activation.

I can identify two positive aspects of using OT, and StOT in particular, to describe the data. First, the need to describe the data in terms of markedness constraints as well as faithfulness examines referential choice from an entirely new perspective. Second, by using StOT to model the referential choice the feasibility of the proposed constraints could be tested by comparing the outputs it would generate with those that appeared in the experiment. This seems to be a good method to test how well a set of constraints accounts for the data.

In the StOT grammar the faithfulness constraints tended to be ranked above all the markedness constraints. There are two possible explanations for this. First, the constraints proposed and the data analyzed are far from complete. In normal writing every sentence does not begin with an adnominal noun phrase, and the elicitation task set-up is therefore unnatural, but necessarily so in order to make a forced comparison. But we do have a richer range of referential forms available including definites and pronominal forms. When these other forms are taken into account, we will necessarily use a larger set of constraints, and then we should expect that markedness constraints play a greater role. Because the data examined were very similar referential expressions that were highly variable, we didn't get to fully see how an OT analysis could contribute to our understanding of referential choice. We need to expand on this work and look at additional referential forms. For this, we need more theoretical proposals like Zeevat's (2002) as well as the corpus study of naturally produced data that would make a necessary complement to the elicitation task.

Second, faithfulness in referential forms, and for demonstratives in particular, may be more important than for other linguistic phenomena. Diessel's (2000) work has shown that from a typological perspective, demonstratives are unusual referential forms in that they exist in all languages, but there is little evidence that they have lexical origins, which grammatical markers are generally believed to develop from. He suggests that this may be because they represent one of the most basic forms present in every language. Perhaps whatever core meaning is coded by demonstratives is so basic that there are no languages that choose not to code it, which means input features associated with demonstratives would be unlikely to be dominated by a markedness constraint.

Additionally, at first glance the results might tempt us to collapse all four markedness constraints into one (and I did just that in one version of the grammar). They do all share one core characteristic: avoid marking that which can be determined from the context. However, this would make it impossible to distinguish markedness violations for one input feature from another, which becomes important when the input has more than one relevant feature.

In summary, this small study seems to show that there are some systematic differences in the choice of the two demonstrative forms, but that this is a subtle difference which needs to be studied more carefully with more data. However, discussing how the results could be modeled in an optimality framework highlighted some issues that need to be addressed in OT pragmatics, namely, the role of markedness in pragmatic phenomena, and the way in which the context should be treated in an OT analysis.

Notes

1. Actually, Swedish can be said to have three different demonstrative forms as it is also possible to use the preposed definite marker with the indefinite noun form,

as in *det hus* and *den hund*, but in such cases the noun phrase is often further modified by a restrictive relative clause.

2. Historically, Old Norse had adnominal demonstratives that followed their nouns. This form then developed into the definite suffix used in modern Swedish.
3. By accessibility I do not mean the logical accessibility as the term is used in DRT.
4. This is because the predicate of the final sentence strongly influences the set of referents that would be natural to refer to in the subject position. While it is possible to refer to entities mentioned in each of the first three sentences, the grammatical role of the antecedents and anchors is then fixed.
5. Both logistic regression and linear regression/ANOVA belong to the family of Generalized Linear Models, but an ANOVA is used with dependent variables that are continuous and assumes a normal distribution, while logistic regression is used with binary dependent variables and assumes a logistic distribution. The dependent variable in this study is discrete and binary so an ANOVA is not appropriate.
6. A step-down analysis was also performed and resulted in the same factors being identified as significant for the model.
7. Beaver's (to appear) OT version of Centering Theory follows standard CT in that it only addresses the choice between names and personal pronouns.
8. Boersma's StOT is very similar to generalized linear models such as logistic regression. For more information on the relationship between them see Paollilo (2002).
9. animate $+ -1$, animate $+ -3$, concrete $+ -1$, concrete $+ -3$, abstract $+ -1$, abstract $+ -3$.
10. First with a plasticity of 2 and an evaluation noise of 10, and then again with a plasticity of 0.2 and an evaluation noise of 2.
11. In fact, I tested at least a dozen, less successful and more or less intuitive constraint sets before settling on the one presented here.

Appendix of stories used in study

Story	*Swedish*	*English translation*
1	Eleverna vantrivs allt oftare i skolan.	*Students are often unhappy in school.*
	Många klasser har alldeles för små klassrum.	*Many classrooms are far too small.*
	Lärarna i vår trångbodda skola är väldigt upprörda.	*Teachers in our cramped school are very upset.*
	C.________ måste byggas ut eller så måste studenterna flyttas.	*C. ________ must be expanded or the students need to be moved.*
	A.________ är inte lätt att uthärda i längden.	*A.________ have/has become unbearable.*
2	Skvaller på arbetsplatsen kan ha en viktig funktion.	*Gossip in the workplace can fill an important function.*
	Det skapar närhet mellan de som skvallrar genom att ge en "vi-mot-dem" känsla.	*It creates a closeness between those who gossip by giving an "us-against-them" feeling.*
	På vår arbetsplats finns det många romantiska intriger bland de anställda.	*At our workplace there are many romantic intrigues between employees.*
	C._______ blir föremål för många spekulationer och diskussioner.	*C._______ have/has become the object of many speculations and discussions.*
	A._______kan bli väldigt generande för de inblandade.	*A. _______ can be very embarrassing for those involved.*
3	Många tycker att det är svårt att hitta rätt bland alla nya produkter som finns och det är därför Modern Teknik har infört en "Nya Produkter"-spalt.	*Many find it difficult to find the appropriate new product among all the new products that come out, and that's why "Modern Teknik" has introduced a "New Products" column.*
	CD-bränning är ett nytt sätt att förvara data på som också har skapat en del förvirring bland konsumenterna.	*CD-burning is a new way to store data that has also created a bit of confusion among consumers.*
	Och affärsbiträdena vet ofta inte särskilt mycket mer än kunderna själva.	*And store workers often also do not know much more than the customers themselves.*
	A. ________ leder ofta till att många köper fel produkt.	*A.__________ often lead(s) to many buying the wrong product.*
	C. _______ beskriver i enkla termer hurman väljer rätt CD-brännare för sina behov.	*C. __________describe(s) in simple terms how to choose the right CD-burner for one's needs.*

Story	*Swedish*	*English translation*
4	Många undersökningar visar en ökning av diabetes, eller sockersjuka, bland kvinnor inom den statliga sektorn.	*Many research studies show an increase in diabetes or "sugar sickness" among women working in the government sector.*
	Stillasittande arbete kombinerat med en brist på fysisk aktivitet leder till en osund livsstil.	*Passive work combined with a lack of physical activity leads to an unhealthy lifestyle.*
	Undersökningar visar också att den sociala delen av kvinnornas arbete ofta går ut på att umgås över ett osunt mellanmål.	*The research also shows that the social part of the women's work often involves chatting over an unhealthy snack.*
	A. ________ förekommer inte bland kvinnor i den privata sektorn.	*A.___________doesn't/don't occur among women in the private sector.*
	C. __________ åt under vissa perioder stora mängder godis.	*C. ________ate during certain periods great amounts of candy.*
5	Kassörskan kommer aldrig ihåg priserna på frukt och grönt.	*The check-out woman never remembers the prices of fruits and vegetables.*
	Affärerna verkar ändra priser hur som helst.	*The stores seem to change the prices willy-nilly.*
	Det händer ibland att kunder blir förargade och de klagar väldigt högljutt.	*It sometimes happens that customers become angry and complain very loudly.*
	A._________ kommer att tas upp på nästa möte hos konsumentverket.	*A. __________ is/are going to be discussed at the next meeting of the consumer department.*
	C._________ skriver brev till butikschefer eller konsumentverket.	*C. __________ write(s) letters to shop managers or to the consumer department.*
6	Många vet inte vilka meriter de ska ta med i sin CV.	*Many do not know what merits they should include in their CV.*
	Att skriva in information om sin utbildning och tidigare jobb är ju självklart.	*Writing information about one's education and earlier jobs is certainly a must.*
	Det är svårare att veta hur mycket man ska ta upp om hobbies och andra fritidsaktiviteter.	*It's much harder to know how much one should bring up about hobbies and free-time activities.*
	A._________ ställer till problem för många som skriver CV för första gången.	*A. _________ make(s) problems for many who write a CV for the first time.*

Story	*Swedish*	*English translation*
	C.________ kan antigen ge ett bra intryck av en välanpassad individ, eller skapa en känsla av trivialitet.	*C. ________ can either give a good impression of a well-adjusted individual or create a feeling of triviality.*
7	Gislaved Energi vill underlätta för sina kunder att betala räkningar.	*Gislaved Energy wants to make it easier for their customers to pay their bills.*
	Därför har vi utvecklat e-räkningar som skickas till kunderna via epost.	*This is why they have developed e-bills that are sent to customers via email.*
	E-räkningar är inte bara praktiskt utan bidrar också till en mindre belastning på miljön tack vare den minskade pappersförbrukningen.	*E-bills are not only practical, but also contribute to a smaller burden for the environment thanks to the decrease in the use of paper.*
	A. _______ visar på framtidstänkande och miljömedvetenhet, vilket kännetecknar Gislaved Energi som bolag.	*A. ________ show(s) a focus on the future and an environmental awareness which characterizes Gislaved Energy as a company.*
	C. _______ är på alla sätt bra för Gislaved Energi's kunder.	*C. ________ is in all ways good for Gislaved Energy's customers.*

11
Learning Constraint Subhierarchies: The Bidirectional Gradual Learning Algorithm

Gerhard Jäger

1 Differential case marking

It is a common feature of many case marking languages that some, but not all objects are case marked.[1] However, it is usually not entirely random which objects are marked and which aren't. Rather, case marking only applies to a morphologically or semantically well-defined class of NPs. Take Hebrew as an example. In this language, definite objects carry an accusative morpheme while indefinite objects are unmarked.

(1) a. Ha-seret her?a *?et-ha-milxama*
THE-MOVIE SHOWED ACC-THE-WAR
b. Ha-seret her?a *(*?et-)milxama*
THE-MOVIE SHOWED (*ACC-)WAR
(from Aissen 2000)

Similar patterns are found in many languages. Bossong (1985) calls this phenomenon "Differential Object Marking" (DOM). A common pattern is that all NPs from the top section of the *definiteness hierarchy* are case marked while those from the bottom section are not:

(2) personal pronoun > proper noun > definite full NP > indefinite specific NP > non-specific indefinite NP

Catalan, for instance, only marks personal pronouns as objects. In Pitjantjatjara (an Australian language), pronouns and proper nouns are case marked when they are objects while other NPs aren't. Hebrew draws the line between definite and indefinite NPs and Turkish between specific and non-specific ones.[2]

Likewise, the criterion for using or omitting a case morpheme for objects may come from the animacy hierarchy:

(3) human > animate > inanimate

As with the definiteness hierarchy, there are languages which only mark objects from some upper segment of this scale. Finally, there are instances of DOM where case marking is restricted to an upper segment of the product of the two scales.[3] Differential case marking also frequently occurs with subjects.[4] In contradistinction to DOM, DSM ("Differential Subject Marking") means that only instances of some *lower* segment of the definiteness/animacy hierarchy are case marked. (The observation that the relevant scales for subjects and objects are inverses of each other is due to Silverstein, 1976.)

DOM and DSM may co-occur within one language. This phenomenon is usually called *split ergativity*. (This term covers both case marking systems where the case marking segments for subjects and for objects are complementary and systems where they overlap.)

The person specification of NPs induces another hierarchy. Simplifying somewhat, it says that the local persons (first and second) outrank third person.

(4) 1st/2nd person > 3rd person

These patterns underlie split ergative case marking in languages like Dyirbal where the choice between the nominative/accusative system and the ergative/absolutive system is based on person. Table 1 (which is taken from Aissen, 1999) shows the basic case marking pattern for Dyirbal.

Briefly put, Dyirbal only marks non-harmonic arguments, that is, local objects and third person subjects. It thus represents a combination of DOM with DSM.

These patterns of "Differential Case Marking" (DCM) can be represented as the result of aligning two scales – the scale of grammatical functions (subject vs. object) with some scale which classifies NPs according to substantive features like definiteness, egocentricity, or animacy (as proposed in Silverstein, 1976). Ranking the grammatical functions according to prominence leads to

Table 1 Case marking system of Dyirbal

	Unmarked	**Marked**
Local persons	Subject	Object
3rd person	Object	Subject (of transitive)
Case	Nominative/Absolutive	Accusative/Ergative

the binary scale:

(5) Subj > Obj

Harmonic alignment of two scales means that items which assume comparable positions in both scales are considered most harmonic. For alignment of the scale above with the definiteness hierarchy this means that pronominal subjects (+ prominent/+ prominent), as well as non-specific objects (− prominent/− prominent) are maximally harmonic, while the combination of a prominent position in one scale with a non-prominent position in the other scale is disharmonic (like non-specific subjects or pronominal objects). More precisely, harmonically aligning the hierarchy of syntactic roles with the definiteness hierarchy leads to two scales of feature combinations, one confined to subjects, and the other to objects. The subject scale is isomorphic to the definiteness hierarchy, while the ordering for objects is reversed:

(6) a. Subj/pronoun > Subj/name > Subj/def > Subj/spec > Subj/non-spec
    b. Obj/non-spec > Obj/spec > Obj/def > Obj/name > Obj/pronoun

In this way DCM can be represented as a uniform phenomenon – case marking is always restricted to upper segments of these scales. This pattern becomes even more obvious if optional case marking is taken into account. As Aissen points out, if case marking is optional for some feature combination, it is optional or obligatory for every feature combination that is lower in the same hierarchy, and it is optional or prohibited for every point higher in the same hierarchy. Furthermore, if one looks at actual frequencies of case marking patterns in corpora, all available evidence suggests that the relative frequency of case marking always increases the farther down one gets in the hierarchy (see Aissen and Bresnan, 2002). What is interesting from a typological perspective is that there are very few attested cases of "inverse DCM" – languages that would restrict case marking to lower segments of the above scales.[5] The restriction to upper segments appears to be a strong universal tendency.

2 OT formalization

Prince and Smolensky (1993) develop a simple method to translate harmony scales into OT constraints: for each element *x* of a scale we have a constraint **x* ("Avoid *x*!"), and the ranking of these constraints is just the reversal of the harmony scale. For the person/grammatical function interaction discussed above, this looks schematically as in (7) (adapted from Bresnan, Dingare and Manning, 2001).

To translate harmony scales into OT, first every feature combination *f* is compiled into a constraint saying "Avoid *f*!" For instance, the combination 'Subj/local' corresponds to the constraint "*Subj/local", that is violated by every local person subject. The ordering in the harmony scale is translated into universal subhierarchies which are to be respected by any language particular total constraint ranking. If, according to the harmony scale, local person subjects are better than third person subjects, then being a third person subject is (universally) worse than being a local person subject. This is expressed by the constraint subhierarchy "*Subj/3rd ≫ *Subj/local":

(7)	Prominence scales	Harmonically aligned scales	OT constraint subhierarchies
	Subj > Obj	Subj/local > Subj/3rd	*Subj/3rd ≫ *Subj/local
	local > 3rd	Obj/3rd > Obj/local	*Obj/local ≫ *Obj/3rd

Generally, the common pattern of DCM is that non-harmonic combinations must be morphologically marked while harmonic combinations are unmarked. To formalize this idea in OT, Aissen employs the formal operation of *constraint conjunction* from Smolensky (1995). If C_1 and C_2 are constraints, C_1 & C_2 is another constraint which is violated iff both C_1 and C_2 are violated. Crucially, C_1 & C_2 may outrank other constraints C_i that in turn outrank both C_1 and C_2. So the following constraint ranking is possible:

$$C_1 \,\&\, C_2 \gg C_3 \gg C_4 \gg C_1 \gg C_5 \gg C_2$$

Furthermore, two general constraints play a role:

- "*Ø" is violated if a morphological feature is not marked
- "*STRUC" is violated by any morphological marking

Each constraint resulting from harmonic alignment is conjoined with *Ø, and the ranking of the conjoined constraints is isomorphic to the ranking induced by alignment. (Also the conjoined constraints outrank each of their conjuncts.) The alignment of the person hierarchy with the scale of grammatical functions thus, for instance, leads to the following universal constraint subhierarchies:

(8) *Ø & *Subj/3rd ≫ *Ø & *Subj/local
*Ø & *Obj/local ≫ *Ø & *Obj/3rd

Interpolating the constraint *STRUC at any point in any linearization of these subhierarchies leads to a pattern where morphological marking indicates non-harmony. The choice of the threshold for morphological marking depends on the relative position of *STRUC. The Dyirbal pattern, for instance, would follow from the following constraint ranking:

(9) *Ø & *Subj/3rd >> *Ø & *Obj/local >> *STRUC >> *Ø & *Subj/local >> *Ø & *Obj/3rd

3 Statistical bias

In Zeevat and Jäger (2002) (ZJ henceforth) we attempt to come up with a functional explanation for the DCM patterns that are analyzed by Aissen. The basis for this approach is the observation that harmonic combinations of substantive and formal features (like the combinations "subject + animate" or "object + inanimate") are common in actual language use, while disharmonic combinations (like "subject + inanimate" or "object + animate") are rather rare. This intuition has been confirmed by several corpus studies. Table 2 displays the relative frequencies of feature combinations in the corpus SAMTAL, a collection of everyday conversations in Swedish that was annotated by Oesten Dahl. (Only subjects and direct objects of transitive clauses are considered.)

There are statistically significant correlations between grammatical function and each of the substantive features definiteness, pronominalization and animacy. The correlations all go in the same direction: harmonic combinations are overrepresented, while disharmonic combinations are underrepresented. If attention is restricted to simple transitive clauses, the chance that an arbitrarily picked NP is a subject is (of course) exactly 50 percent – exactly as high as the chance that it is a direct object. However, if an NP is picked at random and it turns out to be definite, the likelihood that it is a subject increases to 62.9 percent. On the other hand, if it turns out to be indefinite, the probability that it is a subject is as low as 3.9 percent.

Table 2 Frequencies in the SAMTAL corpus of spoken Swedish

	NP	**+def**	**−def**	**+pron**	**−pron**	**+anim**	**−anim**
Subj	3151	3098	53	2984	167	2948	203
Obj	3151	1830	1321	1512	1639	317	2834
χ^2		1496		1681		4399	
$p < 0.01\%$		yes		yes		yes	

Analogous patterns obtain for all combinations:

Table 3 Conditional probabilities

$p(subj \mid +def) = 62.9\%$	$p(subj \mid -def) = 3.9\%$
$p(obj \mid +def) = 37.1\%$	$p(obj \mid -def) = 96.1\%$
$p(subj \mid +pron) = 66.4\%$	$p(subj \mid -pron) = 9.2\%$
$p(obj \mid +pron) = 33.6\%$	$p(obj \mid -pron) = 90.8\%$
$p(subj \mid +anim) = 90.3\%$	$p(subj \mid -anim) = 6.7\%$
$p(obj \mid +anim) = 9.7\%$	$p(obj \mid -anim) = 93.3\%$

This statistical bias has little to do with the grammar of the language at hand. There is some minor influence because diathesis can be used to avoid disharmonic combinations (see Aissen, 1999; and Bresnan, Dingare and Manning, 2001 for discussion), but since the passive is generally rare and there is no categorical grammaticalized correlation between referentiality or animacy and diathesis in Swedish, the general pattern is hardly affected by this factor.[6] So despite the thin cross-linguistic evidence (though the same patterns have been found in the Wall Street Journal Corpus by Henk Zeevat, in the CallHome corpus of spoken Japanese by Fry, 2001, and in the SUSANNE and CHRISTINE corpora of written and spoken English by the present author), I henceforth assume the working hypothesis that these statistical biases are universal features of language use.

4 Bias and bidirectional optimization

Differential case marking amounts to a preference for case marking of disharmonic feature combinations over case marking of harmonic combinations. Taking the statistical patterns of language use into account, this means that there is a preference for case marking of rare combinations, while frequent forms are more likely to be unmarked. This is a sensible strategy because it minimizes the overall effort of the speaker while preserving the disambiguating effect of case marking.[7] As pointed out in ZJ, Bidirectional Optimality Theory in the sense of Blutner (2000) provides a good theoretical framework to formalize this kind of pragmatic reasoning.

According to bidirectional OT (which is founded in work on formal pragmatics), a meaning-form pair is only optimal if it conforms to the preferences of both speaker and hearer in an optimal way. Speaker preferences and hearer preferences of course need not coincide. However, they do not contradict each other either, for the simple reason that the speaker has preferences between different ways to express a given meaning, while the hearer compares different interpretations of a given form. Applied to case marking, it is plausible to assume that the speaker has *ceteris paribus* a preference to avoid case marking. The hearer, on the other hand, has a preference for faithful interpretation (accusative NPs are preferredly interpreted as objects and ergative NPs as subjects). Furthermore, ZJ claim that there is a hearer preference to follow the statistical bias, that is, to interpret definite or animate NPs as subjects and indefinite or inanimate NPs as objects.

These preferences can easily be interpreted as OT constraints. The speaker preference against case marking is just Aissen's constraint *STRUC. Preference for faithfulness interpretation of case morphemes can be covered by a constraint FAITH (arguably there are different faithfulness constraints for different morphemes, but for the purposes of ZJ as well as of the present chapter, one big faithfulness constraint will do). Finally, ZJ assume a constraint BIAS that is fulfilled if an NP of a certain morphological category is interpreted as having the grammatical function that is most probable for this category.[8]

For FAITH to take any effect, it must be (universally) ranked higher than BIAS. The relative ranking of *STRUC is actually inessential. For the sake of illustration, we assume it to be ranked lowest. So the hierarchy of constraints is

FAITH >> BIAS >> *STRUC

In contradistinction to standard OT, bidirectional OT takes both hearer preferences and speaker preferences into account. Hearer optimality means: for a given form, choose the meaning that has the least severe constraint violation pattern. For the constraint system at hand, this means: interpret an NP according to its case marking, and in the absence of case marking, follow the statistical bias. The speaker has to take this hearer strategy into account to get his message across. Only if two competing forms are both hearer optimal for a given meaning, the speaker is free to choose the preferred one (which means in the present set-up: the one without case marking).

This view on bidirectional optimization can be formalized in the following way.[9] I write $\langle m_1, f_1\rangle < \langle \mathrm{m}_2, f_2\rangle$ iff the meaning-form pair $\langle m_1, f_1\rangle$ is better than $\langle m_2, f_2\rangle$ according to the constraints given above in the given ranking. Following standard practice, I assume a generator relation **GEN** between forms and meanings from which the optimal candidates are chosen. **GEN** supplies the morphological inventory of a language as well as some general, highly underspecified structural relation between forms and meanings.

Definition 1

- A meaning–form pair $\langle m, f\rangle$ is *hearer-optimal* iff $\langle m, f\rangle \in$ **GEN** and there is no alternative meaning m' such that $\langle m', f\rangle \in$ **GEN** and $\langle m', f\rangle < \langle m, f\rangle$.

- A meaning–form pair $\langle m, f\rangle$ is *optimal* iff it is hearer-optimal and there is no alternative form f' such that $\langle m, f'\rangle$ is hearer-optimal and $\langle m, f'\rangle < \langle m, f\rangle$.

Now suppose the **GEN** relation for a given language supplies both an accusative and an ergative morpheme. How would, say, an inanimate object be morphologically realized in an optimal way? To keep things simple, let us assume that the interpretation of an NP within a clause is uniquely determined by its grammatical function. (In a more elaborate system, grammatical functions only mediate between surface realization and semantic roles, but I will not go into that in the context of the present chapter.) We get the following tableau:

(10)

		FAITH	BIAS	*STRUC
anim + Ø ☞	Subj			
	Obj		*	
anim + ERG	Subj			*
	Obj	*	*	*
anim + ACC	Subj	*		*
☞	Obj		*	*

To figure out which meaning-form pairs are hearer optimal, we have to compare the different meanings (subject vs. object) of the three potential morphological realizations: zero (i.e., identical to the subject marking in intransitive clauses), ergative or accusative. It is easy to see that the association of both zero marking and ergative marking with the subject role, and the association of accusative marking with the object role are hearer optimal. Speaker optimization chooses between the possible hearer-optimal realizations of a given meaning. For the subject interpretation, there is a choice between zero marking and ergative marking. Since the latter violates *STRUC and the former doesn't, and they do not differ with respect to other constraints, zero marking is optimal for the subject interpretation. For the object interpretation, there is only one hearer optimal realization – accusative marking – which is thus trivially optimal.

For inanimate NPs, the pattern is reversed. Here subjects must be case marked with the ergative morpheme, while objects are preferredly unmarked.

(11)

		FAITH	BIAS	*STRUC
inanim + Ø	Subj		*	
☞	Obj			
inanim + ERG ☞	Subj		*	*
	Obj	*		*
inanim + ACC	Subj	*	*	*
	Obj			*

ZJ's system thus predicts a split ergative system: case marking is restricted to disharmonic feature combinations – animate objects and inanimate subjects – while harmonic combinations are unmarked.

This mechanism only works, though, if the NP at hand is in fact ambiguous between subject and object interpretation. If it is disambiguated by means of external factors like word order, agreement, semantic plausibility and so on, zero marking will always win. Let us call such a case marking system *pragmatic DCM*. However, the languages that were mentioned in the beginning require case marking of disharmonic combinations regardless of the particular contextual setting. Restricting attention to (in)animacy, this would mean that *all* animate objects and inanimate subjects must be case marked, no matter whether case marking is necessary for disambiguation in a particular context. I call such a system *structural DCM* henceforth. Bidirectional OT does not give an immediate explanation for such a pattern.

ZJ suggest that structural DCM emerges out of pragmatic DCM as the result of a grammaticalization process. If a language starts employing pragmatic DCM, the next generation of language learners are faced with two ways of making sense of the case marking pattern: pragmatic DCM or optional structural DCM. If both hypotheses are entertained, the overall probability for DCM increases (i.e., the probability for an animate subject to be zero-marked, for an inanimate subject to be case marked, etc.). This in turn makes the hypothesis of structural DCM more plausible. After some generations of partial reanalysis, DCM is fully grammaticalized, that is, pragmatic DCM has turned into structural DCM.

There are quite a few problems that are left open by the ZJ approach. To start with, the explanation of structural DCM rests on a fairly sketchy account of grammaticalization. Also, it predicts that in languages that have both ergative and accusative morphology at their disposal, a split ergative system should emerge where the split points for DSM and DOM are identical (as in Dyirbal, see above). While this is the common pattern for split ergativity, there are also languages where the segments for subject marking and for object marking overlap. Dixon (1994, p. 86) mentions the example of Cashinawa (a language from Peru), where all pronouns have case marking in object position, and all third person NPs are case marked in subject position. In other words, third person pronouns occur in three forms: unmarked (as subjects of intransitive clauses), ergative case and accusative case. According to the ZJ system, these pronouns should have a bias either towards a subject or an object interpretation, and thus only the interpretation unsupported by this bias should be marked (or, at any rate, this should have been the situation in the pragmatic DCM language from which the structural pattern of Cashinawa emerged). Also, the ZJ system fails to explain the great cross-linguistic diversity of DCM systems. If DCM is directly rooted in a statistical bias which in turn has extra-linguistic sources, one would expect to find not just the same pattern but also the same split across languages.

There is also a conceptual problem with ZJ's approach. The constraint BIAS makes direct reference to the statistics of language use. While it might be plausible that grammatical rules and constraints are induced from frequencies, it seems unlikely that the internalized grammar of a speaker contains a counter that keeps track of the relative frequencies of feature associations, say. In other words, frequencies may help to explain why and how a certain grammar has been learned, but they are not part of this grammar.

In the remainder of this chapter I will outline a theory that remedies the last problem. While the explanation of pragmatic DCM in terms of bidirectional optimization is preserved, the connection between the statistics of language use and the competence grammar of the speakers of a language is established via a learning algorithm, rather than feeding the statistical information directly into the grammar. This approach solves two puzzles: It explains why the constraint subhierarchies that Aissen assumes to be universal are so common without assuming they are intrinsic to UG, and it gives a formal account of the diachronic shift from pragmatic to structural DCM. The cross-linguistic diversity of the possible split points for DCM is not further discussed in this chapter, but it is likely that this problem can be dealt with in this framework as well.

5 Stochastic optimality theory

Aissen (2000) and Aissen and Bresnan (2002) point out that there is not just a universal tendency towards DCM across languages, but that DCM can also be used to describe statistical tendencies within one language that has, in the traditional terminology, optional case marking. In colloquial Japanese, for example, 70 percent of the inanimate subjects, but only 65 percent of the animate subjects are case marked. Conversely, 54 percent of the animate, but only 47 percent of the inanimate objects are marked (these figures are taken from Aissen and Bresnan, 2002, who attribute them to Fry, 2001). Structural DCM can actually be seen as the extreme borderline case where these probabilities are either 100 percent or 0 percent. Stochastic Optimality Theory (StOT henceforth) in the sense of Boersma (1998) is a theoretical framework that is well-suited to formalize this kind of intuition. As a stochastic grammar, a StOT Grammar does not just distinguish between grammatical and ungrammatical signs, but it defines a probability distribution over some domain of potential signs (in the context of OT: **GEN**). Ungrammaticality is thus the borderline case where the grammar assigns the probability 0.

StOT deviates from standard OT in two ways:

- **Constraint ranking on a continuous scale**: Every constraint is assigned a real number. This number does not only determine the ranking of the constraints, but it is also a measure for the distance between them.
- **Stochastic evaluation**: At each evaluation, the placement of a constraint is modified by adding a normally distributed noise value. The ordering of the constraint after adding this noise value determines the actual evaluation of the candidate set at hand.

So we have to distinguish between the value that the grammar assigns to a constraint, and its actual ranking during the evaluation of a particular candidate. To make this point clear, suppose we have some constraint C which, according to the grammar, has the value 0.5.[10] To evaluate whether a particular linguistic item in a corpus violates this constraint, we have to determine C's actual value. It is obtained from its grammar value by adding some amount z of unpredictable noise. z may be any real number, so the actual value of C can be any number as well. However, z is likely to have a small absolute value, so the actual value of C is likely to be in the vicinity of 0.5. Boersma assumes that z is distributed according to a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 2$.[11] So the actual value of C is also normally distributed, with mean 0.5 and standard Deviation 2. This distribution is depicted in Figure 1.

It has the Gaussian bell shape that is familiar from many stochastic phenomena. The probability that the value of C falls within a certain interval on the x-axis is proportional to the area between the x-axis and the bell curve over this interval. The entire area under the curve is 100 percent. So, for instance, the probability that the value of C is less than 0.5 is exactly 50 percent. While the curve never touches the x-axis on either side in theory, the probability that C is ranked below -9 or over 10 is so small (about 10^{-6}) that it can be ignored.

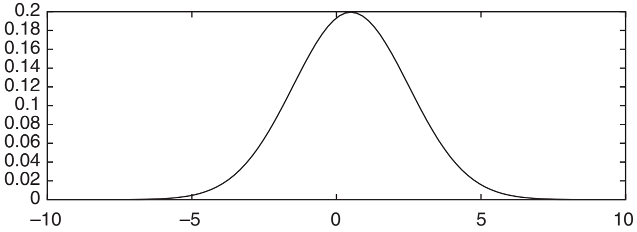


Figure 1 Normal distribution

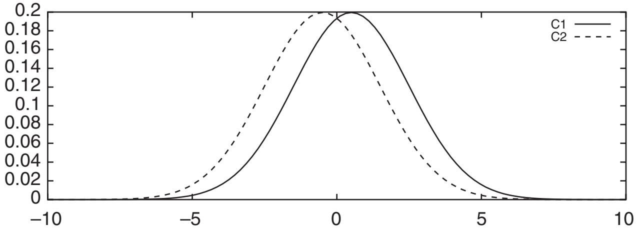


Figure 2 Two constraints

An OT system consists of several constraints, and the addition of a noise value is done for each constraint separately. Suppose the grammar assigns the constraints C1 and C2 the mean values -0.5 and 0.5 respectively. The corresponding function graph is depicted in Figure 2.

Since the mean of C1 is higher than the mean of C2, most of the time C1 will end up having a higher value than C2. However, it is perfectly possible that C2 receives an unusually high and C1 an unusually low value, so that in the end $C2 > C1$.[12] The probability for this is about 36 percent. In any event, after adding the noise values, the actual values of the constraints define a total ranking. This generalizes to systems with more than two constraints. This total ordering of constraints is then used to evaluate candidates in the standard OT fashion, that is, the strongest constraint is used first as a decision criterion; if there is a draw, resort is taken to the second highest constraint and so on. To take the example above, suppose there are two candidates, and the first violates only C1 and the second only C2. In 64 percent of all cases, $C1 > C2$, and thus the first candidate will be selected as optimal. However, in 36 percent of all evaluation events, $C2 > C1$ and thus the second candidate wins.

The probability for $C1 > C2$ depends on the difference between their mean values that are assigned by the grammar. Let us denote the mean values of C1 and C2 as c1 and c2 respectively. Then the probability that C1 outranks C2 is a monotonic function of the difference between their mean values, $c1-c2$.[13] It is depicted in Figure 3.

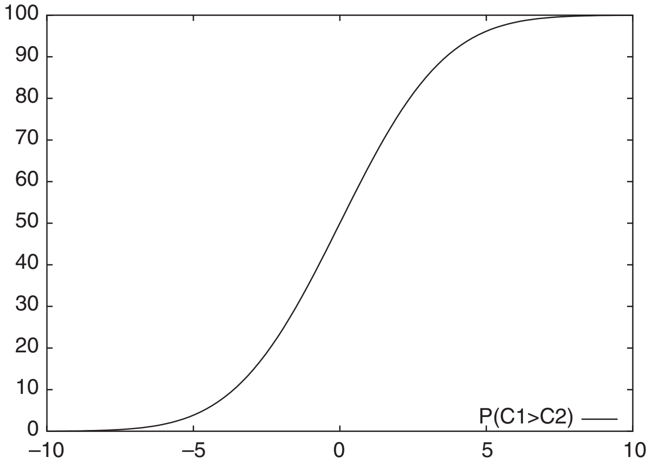


Figure 3 Probability of C1 > C2 as a function of c1−c2 in percentage

If c1 = c2, both have the same chance to outrank the other, and accordingly P(C1 > C2) = 50 percent. This corresponds to a scenario where there is free variation between the candidates favored by C1 and those favored by C2. If C1 is higher ranked than C2, there is a preference for the C1 candidates. If the difference is 2, say, the probability that C1 outranks C2 is already 76 percent. A difference of 5 units corresponds to a chance of 96 percent that C1 > C2. Candidates that are favored by C2 are a rare exception in a language described by such a grammar, but they are still possible. If the difference is larger than 12 units, the probability that C2 outranks C1 is less than 10^{-5}, which means that it is impossible for all practical purposes. In such a grammar C1 always outranks C2, and candidates that fulfill C2 at the expense of violating C1 can be regarded simply as ungrammatical (provided there are alternative candidates fulfilling C1, that is). So the classical pattern of a categorical constraint ranking is the borderline case of the stochastic evaluation. It obtains if the distances between the constraints are suffciently large.

6 The Gradual Learning Algorithm

StOT is equipped with a learning algorithm that extracts a constraint ranking from a representative sample of a language – Boersma's Gradual Learning Algorithm (GLA; see Boersma, 1998; Boersma and Hayes, 2001). A note of caution is in order here: the algorithm only learns a constraint ranking. Both **GEN** and the inventory of constraints have to be known in advance. Furthermore, the algorithm requires as input an analyzed corpus, that is, a set of input-output pairs. (These are pairs of phonological and phonetic representations in the realm of phonology where this system was originally developed. In the present context this amounts to meaning–form pairs.) At every stage of the learning process, the algorithm has its own hypothetical StOT grammar. When it is confronted with an observation, it generates an output for the observed input according to its current grammar and compares it to the observed output. If the two outputs coincide, the observation is taken as confirmation of the hypothetical grammar and no action is taken. If there is a mismatch though, the constraints of the learner's grammar are reranked in such a way that the observed output becomes more likely and the output that the learner produced on the basis of its hypothetical grammar becomes less likely. This process is repeated until further observations do not lead to significant changes of the learner's grammar anymore. If the training corpus is a (sufficiently large) representative sample of a language that was generated by a StOT grammar *G* (which is based on the same **GEN** and constraint set that the learner assumes), the grammar to which the algorithm converges describes a language that assigns the same probabilities to all candidates as *G*. So the learned grammar reproduces the statistical patterns from the training corpus, not just the categorical distinctions between grammatical and ungrammatical. Note that the algorithm is *error-driven* – the learner revises his hypothesized grammar only if there is a discrepancy between the observations and her own preferences.

Schematically, the algorithm goes through six different stages during the learning process:

- **Initial state** All constraint values are set to 0.
- **Step 1: a datum** The algorithm is presented with a learning datum – a fully specified input-output pair $\langle i, o\rangle$.
- **Step 2: generation**
  - For each constraint, a noise value is drawn from a normal distribution and added to its current ranking. This yields the *selection point*.
  - Constraints are ranked by descending order of the selection points. This yields a linear order of the constraints.
  - Based on this constraint ranking, the grammar generates an output o' for the input i.
- **Step 3: comparison** If $o = o'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle i, o\rangle$ with the self-generated pair $\langle i, o'\rangle$.
- **Step 4: adjustment**
  - All constraints that favor $\langle i, o\rangle$ over $\langle i, o'\rangle$ are *increased* by some small predefined numerical amount ('plasticity').

  - All constraints that favor ⟨*i, o′*⟩ over ⟨*i, o*⟩ are *decreased* by the plasticity value.

- **Final state** Steps 1–4 are repeated until the constraint values stabilize.

There are several numerical parameters involved that influence the behavior of the GLA to a certain degree. I assume here that all constraints start with the initial value 0 (Boersma and Hayes, 2001, use 100 here). The concrete value is totally inessential. The plasticity value is crucial for the impact of a single observation and thus for the speed of learning. A high plasticity accelerates learning at the expense of allowing a single observation to have a high impact. Conversely, a low plasticity makes the algorithm slower but more robust.[14] In the context of the present chapter, I will use a plasticity of 0.01.

7 GLA and grammaticalization

Cable (2002) makes an ingenious proposal regarding how the GLA can be used to explain the shift from pragmatic to structural DCM – a problem that was largely left open by ZJ. Suppose a language has reached a stage where pragmatic DCM is mandatory, while there is no structural DCM (yet). Let us also assume, for the sake of the argument, that the relative frequencies are as in the SAMTAL corpus (see Figure 2), and that on the average one out of two NPs is unambiguous with respect to its grammatical role (for instance due to word order). We restrict attention to the contrast +/−animacy. (Cable's example is the contrast between local persons and third person, which makes no difference to the argument.) In this language, there is never case marking on animate subjects or inanimate objects because such NPs are either unambiguous to start with, or if they are ambiguous, BIAS assigns them the correct interpretation. Disharmonic combinations (inanimate subjects and animate objects) are marked whenever they are otherwise ambiguous, that is, in 50 percent of all cases by assumption. Now suppose this language is fed into the GLA based on a **GEN** that supplies both ergative and accusative morphology. The constraints to be ranked are Aissen's: *Subj/anim & *Ø, *Subj/inanim & *Ø, *Obj/anim & *Ø, *Obj/inanim & *Ø, and *STRUC. Since in the language under discussion 50 percent of all inanimate subjects are case marked, the GLA converges to a ranking where *Subj/inanim & *Ø and *STRUC have the same rank (and thus their two possible rankings with respect to each other are equally likely, leading to a 50 percent preference in favor and a 50 percent preference against ergative marking of inanimate subjects). The same applies to animate objects. Animate subjects and inanimate objects are never case marked, so the constraints *Subj/anim & *Ø and *Obj/inanim & *Ø end up being ranked well below *STRUC so that it is virtually impossible for them to outrank *STRUC.

A ranking with these properties would be:

(12) a. *STRUC = *Subj/inanim & *Ø = *Obj/anim & *Ø = 5
b. *Subj/anim & *Ø = *Obj/inanim & *Ø = −5

The next generation uses this grammar but also employs pragmatic DCM for disambiguation. Hence it will also never use case marking for harmonic combinations – neither the grammar nor pragmatics gives a reason to do so. Now the grammar requires that 50 percent of all disharmonic NPs are case marked, but the correlation between case marking and ambiguity is lost. On average, half of the ambiguous and half of the unambiguous disharmonic NPs are marked for grammatical reasons. If a disharmonic NP is per se ambiguous and happens to be unmarked by the grammar, the pragmatic DCM strategy requires it to be marked nevertheless. Hence, in the end this generation will use case marking for 75 percent of all disharmonic NPs. The next generation will thus learn a grammar where *Subj/inanim & *Ø and *Obj/anim & *Ø are placed 2 units higher than *STRUC to mimic this 75 : 25 proportion, while *Subj/anim & *Ø = *Obj/inanim & *Ø are again way below *STRUC. The grammar that is learned by this generation looks like:

(13) a. *Subj/inanim & *Ø = *Obj/anim & *Ø = 7
b. *STRUC = 5
c. *Subj/anim & *Ø = *Obj/inanim & *Ø = −5

By the same kind of reasoning, in the next generation this ratio will rise to 87.5 percent and so on. After ten generations 99.9 percent of all disharmonic NPs, but still none of the harmonic NPs, are case marked. In other words, pragmatic DCM has turned into structural DCM.

8 Learning and bidirectionality

Cable's approach solves one big problem that ZJ leave open: it describes a precise mechanism of grammaticalization of DCM, the shift from pragmatic towards structural DCM. The other big problem is still open: the whole mechanism is driven by pragmatic DCM which in turn is based on the constraint BIAS, and thus mixes grammar with statistical tendencies. Also, Cable's mechanism in a sense assumes that the learner is pragmatically ignorant – it is confronted with pragmatic DCM and mistakenly analyzes it as optional structural DCM. After completion of learning, however, the next generation reinvents pragmatic DCM on top of the acquired structural DCM. So the learner uses a different type of grammar than the adult speaker.

These shortcomings can be overcome by extending bidirectional optimization to the learning process. Assume that the training corpus is drawn from a language that was generated by a StOT grammar based on bidirectional optimization in the sense of Definition 1. (As argued in the discussion of ZJ above, this has the advantage that pragmatic DCM is integrated into the OT machinery.) Accordingly, the same bidirectional notion of optimality should be used by the learning algorithm in the second step (generation). Recall that the learner takes the observed input and generates an output for that input on the basis on her current hypothesized grammar. This output has to be optimal on the basis of the hypothesized grammar, and in the bidirectional version of the GLA, 'optimal' means 'bidirectionally optimal'.

There is a minor problem with this adjustment though. For the generation step of the GLA to succeed, it has to be guaranteed that there is some optimal output for each observed input. This is always the case according to standard (unidirectional) optimization, but it need not be the case if one uses bidirectional optimization in the sense defined above.[15] To remedy this, the definition of bidirectional optimality has to be modified somewhat. In its present form, a form is optimal for a given meaning if it is the best option among the hearer-optimal forms for this meaning. We have to add the clause that the optimal form should be the best hearer-optimal form *if there is any*. If no possible form for a given meaning is hearer optimal, we ignore the requirement of hearer optimality. Formally, this reads as:

Definition 2

- A meaning-form pair $\langle m, f \rangle$ is *hearer-optimal* iff $\langle m, f \rangle \in$ **GEN** and there is no alternative meaning m' such that $\langle m', f \rangle \in$ **GEN** and $\langle m', f \rangle < \langle m, f \rangle$.
- A meaning-form pair $\langle m, f \rangle$ is *optimal* iff either it is hearer-optimal and there is no alternative form f' such that $\langle m, f' \rangle$ is hearer-optimal and $\langle m, f' \rangle < \langle m, f \rangle$, or there is no hearer-optimal $\langle m, f' \rangle$, and there is no $\langle m, f' \rangle \in$ **GEN** such that $\langle m, f' \rangle < \langle m, f \rangle$.

You can think of the requirement of hearer-optimality as another constraint that outranks all other constraints. If it is possible to fulfill it, the optimal candidate must do so, but if it cannot be fulfilled it is simply ignored.[16]

Using this notion of optimality together with the GLA, learning involves interpretation as well as generation. This idea of bidirectional learning can be pushed even further by assuming that the learner assumes the hearer perspective and the speaker perspective simultaneously. In Boersma's version of the GLA, the learner observes a datum $\langle m, f \rangle$, generates a pair $\langle m, f' \rangle$ which is optimal according to her own grammar, and then compares f with f'. Bidirectional learning means that the learner also interprets f according to her own grammar and compares the result with the observation.[17] Formally, the learner generates a pair $\langle m', f \rangle$ which is optimal according to her own grammar, and compares m with m'. The next steps – comparison and adjustment – are applied both to m/m' and f/f'. So the bidirectional version of the GLA – call it "Bidirectional Gradual Learning Algorithm" (BiGLA) – is as follows:

- **Initial state** All constraint values are set to 0.
- **Step 1: a datum** The algorithm is presented with a learning datum – a fully specified input-output pair $\langle m, f \rangle$.

- **Step 2: generation**
  - For each constraint, a noise value is drawn from a normal distribution and added to its current ranking. This yields the *selection point*.
  - Constraints are ranked by descending order of the selection points. This yields a linear order of the constraints.
  - Based on this constraint ranking, the grammar generates two pairs $\langle m, f' \rangle$ and $\langle m', f \rangle$ that are both bidirectionally optimal.
- **Step 3.1: comparison of forms** If $f = f'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle m, f \rangle$ with the self-generated pair $\langle m, f' \rangle$.
- **Step 3.2: comparison of meanings** If $m = m'$, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum $\langle m, f \rangle$ with the self-generated pair $\langle m', f \rangle$.
- **Step 4: adjustment**
  - All constraints that favor $\langle m, f \rangle$ over $\langle m, f' \rangle$ are *increased* by the plasticity value.
  - All constraints that favor $\langle m, f' \rangle$ over $\langle m, f \rangle$ are *decreased* by the plasticity value.
  - All constraints that favor $\langle m, f \rangle$ over $\langle m', f \rangle$ are *increased* by the plasticity value.
  - All constraints that favor $\langle m', f \rangle$ over $\langle m, f \rangle$ are *decreased* by the plasticity value.
- **Final state** Steps 1–4 are repeated until the constraint values stabilize.

9 BiGLA and DCM

In this section I will argue that the BiGLA combines the advantages of the ZJ approach to pragmatic DCM and of Cable's theory of grammaticalization. To see this point, let us do a thought experiment. Suppose the BiGLA is confronted with a language that

- has the same frequency distribution of the possible combinations of subject versus object with animate versus inanimate as the spoken Swedish from the SAMTAL corpus,
- always respects FAITH, and
- uses case marking in exactly 50 percent of all cases, but in a way that is totally uncorrelated to animacy. For each clause type, in 25 percent of all cases no case marking is used, in 25 percent the subject is ergative marked and the object is unmarked, in 25 percent the subject is unmarked and the object accusative marked, and in 25 percent both NPs are case marked.

We only consider simple transitive clauses, and we assume that this toy language has no other means for disambiguation besides case marking. So a learning datum will always be a combination of two NPs with a transitive verb. (I also assume that there are no verb specific preferences for certain readings of morphological markings.) Let us call the first NP "NP1" and the second one "NP2".

To see how BiGLA reacts to this language, we have to specify **GEN** and a set of constraints. Strictly speaking, animacy plays a double function in this experiment: it is of course an aspect of the meaning of an NP, but I also assume that this specification for +anim or −anim can be read off directly from the form of an NP. So +anim and −anim are treated as formal features, and **GEN** only relates animate meanings to +anim forms and inanimate meanings to −anim forms. There are thus eight possible semantic clause types to be distinguished because NP1 can be subject and NP2 object or vice versa, and both subject and object can be either animate or inanimate.

Let us assume that **GEN** supplies both ergative and accusative morphology, which are both optional. The linking of case morphemes to grammatical functions is governed by the FAITH constraint, so **GEN** imposes no restrictions in this respect. **GEN** thus admits nine types of morphological marking within a clause: both NP1 and NP2 can be ergative marked, accusative marked or unmarked. This gives nine different form patterns. If +/−anim is taken into account, we get 36 different forms in total. However, **GEN** is organized in such a way that the animacy specification of the forms is completely determined by the meaning. So altogether we end up with 72 meaning–form combinations that are consistent with this **GEN.**

As mentioned above, we extract the frequencies of the possible meanings from the SAMTAL corpus. The absolute numbers are given in Table 4. Not surprisingly, the combination where both subject and object are harmonic is by far the most frequent pattern, and the combination of two disharmonic NPs is very rare.

Table 5 gives a frequency distribution (in percentage of all clauses in the corpus) over this **GEN** which respects the relative frequencies of the different meanings from SAMTAL and treats the linking of NP1 or NP2 to the subject role as equally likely. The notation 'case1-case2' indicates that NP1

Table 4 Frequencies of clause types in SAMTAL

	subj/anim	**subj/inanim**
obj/anim	300	17
obj/inanim	2648	186

Table 5 Training corpus

	E–E	**E–A**	**E–Z**	**A–E**	**A–A**	**A–Z**	**Z–E**	**Z–A**	**Z–Z**
su/a-ob/a	0.0	1.19	1.19	0.0	0.0	0.0	0.0	1.19	1.19
su/a-ob/i	0.0	10.50	10.50	0.0	0.0	0.0	0.0	10.50	10.50
su/i-ob/a	0.0	0.07	0.07	0.0	0.0	0.0	0.0	0.07	0.07
su/i-ob/i	0.0	0.74	0.74	0.0	0.0	0.0	0.0	0.74	0.74
ob/a-su/a	0.0	0.0	0.0	1.19	0.0	1.19	1.19	0.0	1.19
ob/a-su/i	0.0	0.0	0.0	0.07	0.0	0.07	0.07	0.0	0.07
ob/i-su/a	0.0	0.0	0.0	10.50	0.0	10.50	10.50	0.0	10.50
ob/i-su/i	0.0	0.0	0.0	0.74	0.0	0.74	0.74	0.0	0.74

is marked with case1 and NP2 with case2 (E, A and Z abbreviate "ergative", "accusative" and "zero" respectively). Likewise, the notation "su/a-ob/i" means that NP1 is interpreted as animate subject and NP2 as inanimate object and so on.

As for the constraint inventory, I basically assume the system from Aissen (2000) (restricted to the animate/inanimate contrast). This means we have four marking constraints. Using the same notation as in the table above, we can write them as *(su/a/z), *(su/i/z), *(ob/a/z) and *(ob/i/z). They all enforce case marking. They are counteracted by *STRUC which is violated by a clause as often as there are case morphemes present in a clause. (The evaluation of the constraints is done per clause, not just per NP.) The constraint FAITH takes care of the linking between case morphemes and grammatical roles. It is always violated if an ergative marked NP is interpreted as an object or an accusative NP as a subject. Finally, I assume that the grammar does distinguish between interpreting NP1 or NP2 as a subject. In real languages there are many constraints involved here (pertaining to syntax, prosody and information structure). In the context of our experiment, I skip over these details by assuming just two more constraints, SO and OS. They are violated if NP2 is subject and if NP1 is subject respectively. Since all constraints start off with the initial value 0, there is no a priori preference for a certain linking – these two constraints simply equip UG with means to distinguish between the two possible linking patterns. Altogether we thus get eight constraints:

1. *(su/a/z): *Avoid unmarked animate subjects!*
2. *(su/i/z): *Avoid unmarked inanimate subjects!*
3. *(ob/a/z): *Avoid unmarked animate objects!*
4. *(ob/i/z): *Avoid unmarked inanimate objects!*
5. *STRUC: *Avoid case marking!*
6. FAITH: *Avoid ergative marked objects and accusative marked subjects!*
7. SO: *NP1 is subject and NP2 object.*
8. OS: *NP2 is subject and NP1 object.*

All these constraints are set to the initial value 0 and the BiGLA is applied to a training corpus with the frequencies as in Table 5. What is the learning effect of the different observations? Suppose the algorithm is confronted with a clause containing an ergative marked animate subject. In speaker mode, the algorithm produces its own form for the observed meaning (su/a), which may be either ergative marking as well, or else accusative marking or zero marking. The constraint violation profiles of the three candidates at hand are given in (14). (For simplicity, I leave out the last two constraints. The horizontal ordering of the constraints is arbitrary and should not be interpreted as a ranking.)

(14)

	*(su/a/z)	*(su/i/z)	*(ob/a/z)	*(ob/i/z)	*STRUC	FAITH
su/a/erg					*	
su/a/acc					*	*
su/a/z	*					

If the learner's form coincides with the observation, nothing happens. Otherwise, all constraints that favor the observation over the learner's output will be promoted, and all constraints that favor the learner's hypothesis will be demoted. If the learner chooses accusative as its own hypothesis, there is only one constraint that distinguishes between observation and hypothesis, namely FAITH. It favors the observation over the hypothesis and is thus promoted in this scenario. If the learner chooses zero marking, *(su/a/z) favors the observation and is thus promoted, while *STRUC favors the hypothesis and is demoted. The effect of observing other case marked NPs is analogous. So the net effect of observing case marked NPs under the speaker perspective is:

- promotion of *(su/a/z), *(su/i/z), *(ob/a/z), *(ob/i/z) and FAITH
- demotion of *STRUC

Observing unmarked NPs has, by and large, the opposite effect. If an animate subject with zero marking is observed, a mismatch can occur if the learner produces accusative marking or ergative marking. Both will cause a promotion of *STRUC and a demotion of *(su/a/z). In the former case, we will additionally get a promotion of FAITH. The same applies *mutatis mutandis* for other unmarked NPs. So in general, observing unmarked NPs in speaker mode has the following total learning effect:

- promotion of *STRUC and FAITH
- demotion of *(su/a/z), *(su/i/z), *(ob/a/z) and *(ob/i/z)

Since there is the same number of marked and unmarked NPs in the training corpus, we expect these competing forces to cancel each other out, with the exception of FAITH, which is always promoted. So the unidirectional GLA would come up with a grammar where FAITH is high and all other constraints remain around the initial value 0.[18] Such a grammar would reproduce the distribution from the training corpus, that is, 50 percent case marking respecting FAITH but uncorrelated to animacy.

However, the BiGLA also learns in hearer mode, and here the effect is different. First, consider what happens if a case marked NP is observed, for instance an ergative marked animate subject. The possible interpretations are animate subject and animate object. The pattern of constraint violations of the relevant candidates is given in (15):

(15)

	*(su/a/z)	*(su/i/z)	*(ob/a/z)	*(ob/i/z)	*STRUC	FAITH
su/a/erg					*	
ob/a/erg					*	*

The latter but not the former violates FAITH. Due to the effect of speaker learning, FAITH quickly becomes the strongest constraint, so the learner will rarely, if ever, come up with a non-faithful interpretation for an observed form. Hence mismatches between observations and the learner's interpretation are rare. Case marked NPs thus have almost no learning effect in hearer mode.

This is dramatically different for unmarked NPs. Suppose the learner is confronted with an unmarked animate *subject*:

(16)

	*(su/a/z)	*(su/i/z)	*(ob/a/z)	*(ob/i/z)	*STRUC	FAITH
su/a/z	*					
ob/a/z			*			

Now both interpretations, as subject and as object, are consistent with FAITH. So an object interpretation and thus a mismatch is possible. This will lead to a promotion of *(ob/a/z) and a demotion of *(su/a/z). Observing an animate *object*, a mismatch has the opposite effect – promotion of *(su/a/z) and demotion of *(ob/a/z). There are about nine times as many animate subjects as animate objects in the training corpus though. So the net effect of observing unmarked animate NPs is a promotion of *(ob/a/z) and a demotion of *(su/a/z). For inanimate NPs this is exactly the other way round. Here the objects roughly outnumber the subjects by a factor of 14. Hence in total *(su/i/z) will be promoted and *(ob/i/z) demoted.

To summarize, the net effect of learning in hearer mode is:

- promotion of *(su/i/z) and *(ob/a/z)
- demotion of *(su/a/z) and *(ob/i/z)

So bidirectional learning has the effect that the asymmetries in the frequencies of NP types in the training corpus lead to an asymmetric ranking of the corresponding constraints in the learned grammar. Note that Aissen's subhierarchies are in fact induced from the statistics of language use here: *(su/i/z) $\gg$ *(su/a/z), and *(ob/a/z) $\gg$ *(ob/i/z).

A computer simulation revealed that the above considerations are largely correct (except that there is a net demotion of *STRUC). The BiGLA was fed with 50,000 observations which were drawn at random from a distribution as in Table 5. The constraint rankings that were acquired are given in Table 6.

The development of the rankings of the constraints are plotted in Figure 4. The *x*-axis gives the number of observations (in thousands) and the *y*-axis the ranking of the constraints.

In this grammar, FAITH is by far the strongest constraint. Hence the language described by this grammar never uses case marking in an unfaithful way. Further, the disharmonic constraints *(su/i/z) and *(ob/a/z) are ranked well above *STRUC. So case marking of disharmonic NPs is strongly

Table 6 Grammar that was acquired by the BiGLA from the corpus with random case marking

*(su/a/z)	−1.32
*(su/i/z)	2.89
*(ob/a/z)	0.92
*(ob/i/z)	−1.07
*STRUC	−1.05
FAITH	7.94
OS	−0.03
SO	0.03

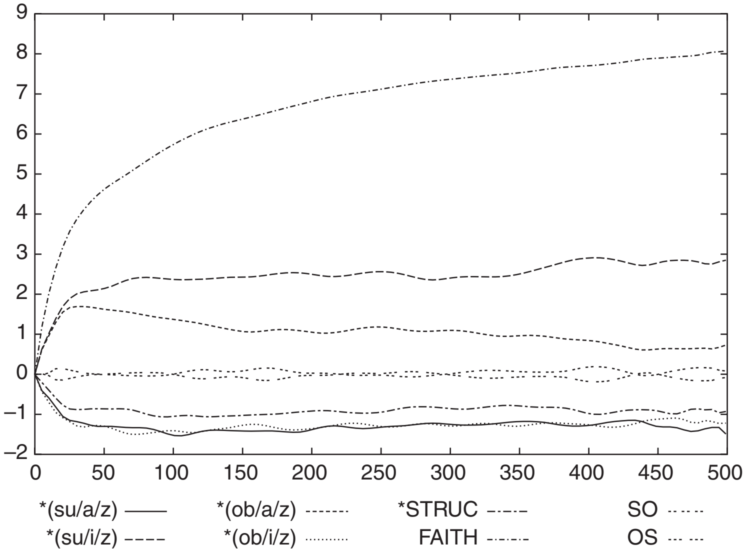


Figure 4 Learning curves

preferred (the distance between the relevant competing constraints is about 4.0 and 2.0 units respectively, which corresponds to a strong preference, but not a categorical rule). The harmonic constraints *(su/a/z) and *(ob/i/z) have about the same ranking as *STRUC – case marking of harmonic NPs is thus totally optional.

These considerations apply if an NP is unambiguous. For an ambiguous unmarked NP, the harmonic interpretation is always preferred because *(ob/a/z) ≫ *(su/a/z) and *(su/i/z) ≫ *(ob/i/z). To achieve bidirectional optimality, this tendency has to be counteracted by using case marking for disharmonic NPs, while harmonic NPs also receive the correct interpretation without case marking. Hence, on top of the preference for structural DCM, there is an even stronger tendency for pragmatic DCM.

The chart in Table 7 below gives the relative frequencies of NP types in a corpus that was generated by maintaining the proportions of meanings from the SAMTAL corpus but using the grammar from Table 6.

Let us consider all cells where the object is accusative marked and the subject is thus not in danger of being understood as an object. Ergative marking is redundant. It is nevertheless used in 60.6 percent of all cases. These 60.6 percent are not equally distributed over animate and inanimate subjects; 95.7 percent of all (unambiguous) inanimate subjects, but only 58.3 percent of all (unambiguous) animate subjects carry ergative case. The same pattern can be observed for objects. Redundant accusative marking is used in 65.2 percent of all cases. However, 83.0 percent of the animate

Table 7 Corpus that was generated by the acquired grammar

	E-E	E-A	E-Z	A-E	A-A	A-Z	Z-E	Z-A	Z-Z
su/a-ob/a	0.0	1.59	0.43	0.0	0.0	0.0	0.0	2.17	0.57
su/a-ob/i	0.0	12.09	7.05	0.0	0.0	0.0	0.0	8.68	13.65
su/i-ob/a	0.0	0.16	0.17	0.0	0.0	0.0	0.0	0.03	0.0
su/i-ob/i	0.0	1.41	1.29	0.0	0.0	0.0	0.0	0.48	0.21
ob/a-su/a	0.0	0.0	0.0	2.08	0.0	1.92	0.29	0.0	0.48
ob/a-su/i	0.0	0.0	0.0	0.29	0.0	0.0	0.79	0.0	0.0
ob/i-su/a	0.0	0.0	0.0	13.49	0.0	8.68	7.12	0.0	12.98
ob/i-su/i	0.0	0.0	0.0	1.32	0.0	0.63	1.46	0.0	0.11

objects, but only 63.3 percent of the inanimate objects are accusative marked (if they co-occur with an ergative marked subject). So we, in fact, see a clear preference for structural DCM.

This effect is more dramatic if we consider potentially ambiguous NPs. In total, 38.9 percent of all subjects that co-occur with an unmarked object are ergative marked. For animate subjects, this figure is 34.8 percent, but for inanimate subjects it is 90.4 percent. As for the objects, 43.6 percent of objects in a clause with an unmarked subject are accusative marked. For animate objects, this figure rises to 79.8 percent, while for inanimate objects it is only 39.4 percent. Of course case marking of subjects and objects influence each other: for the most harmonic meaning (animate subject and inanimate object) 31.5 percent of all clauses use no case marking at all, while for the least harmonic meaning (inanimate subject and animate object) case marking is 100 percent obligatory, only the choice between subject marking, object marking, or both is optional. So, in sum, we see that pragmatic DCM is also present on top of structural DCM.

10 The next generation

The sample corpus that was generated with the acquired grammar can of course be used as input to a second run of the BiGLA. This procedure may be repeated over several "generations". In this way the BiGLA can be used to simulate diachronic development. The successive constraint rankings

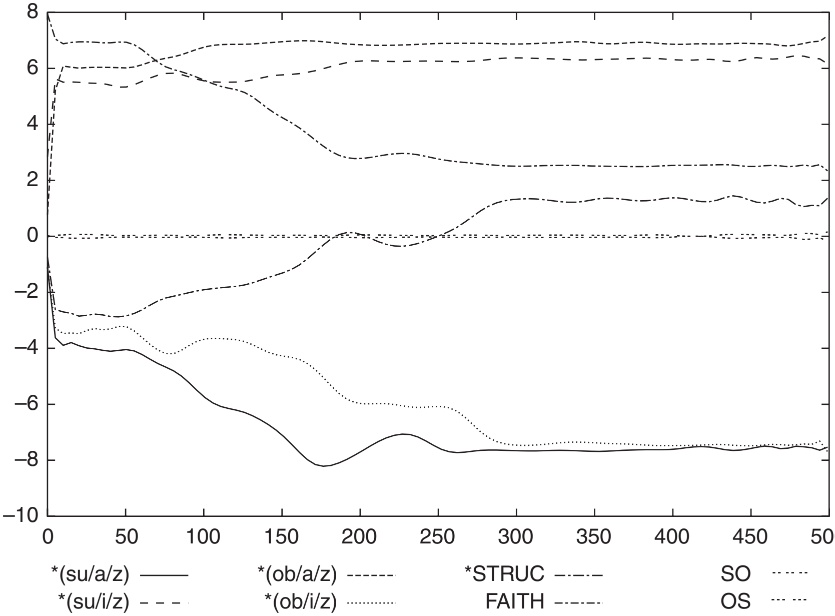


Figure 5 Diachronic development

that emerge in this way are plotted in Figure 5. The learning procedure was repeated 500 times, and the generations are mapped to the *x*-axis, while the *y*-axis again gives the constraint rankings.

While there are no rough changes from one generations to the next, the grammar as a whole gradually changes its characteristics over time. The Aissen subhierarchies – *(su/i/z) ≫ *(su/a/z) and *(ob/a/z) ≫ *(ob/i/z) – are always respected though.

We may distinguish four phases. During the first phase (generations 1–10), the constraints *(su/i/z) and *(ob/a/z) stay closely together, and they increase their distance from *STRUC. This amounts to an ever stronger tendency for case marking of disharmonic NPs. Simultaneously, *(su/a/z) and *(ob/i/z) stay close to *STRUC, that is, we have optional case marking of harmonic NPs. This corresponds to a split ergative system with optional marking of harmonic and obligatory marking of disharmonic NPs. These characteristics remains relatively stable during the second phase (roughly generations 11–60). Then the system becomes unstable. The two constraints pertaining to the disharmonic combinations – *(su/i/z) and *(ob/a/z) – remain high. However, the two "harmonic" constraints *(su/a/z) and *(ob/i/z) are gradually lowered while *STRUC rises. During this process, *STRUC assumes a position strictly below the disharmonic, but strictly above the harmonic case marking constraints. This amounts to a gradual loss of case marking of harmonic NPs, while marking of disharmonic NPs remains obligatory. At around generation 280 this process is completed, and in the remaining 220 generations the system remains stable in a state where case marking is obligatory for disharmonic and prohibited for harmonic NPs. An almost[19] categorical split ergative system has emerged.

The development of the probabilities for of structural (i.e., redundant) case marking of an NP of a given semantic type are given in the left graphics of Figure 6. There the gradual loss of case morphology of harmonic NPs is easy to discern. Needless to say, the diachronic development that is predicted by the BiGLA (together with **GEN**, the constraint set, and the probability distribution over meanings from SAMTAL) depends on the pattern of case marking that was used in the first training corpus. A full understanding of the dynamics of this system and the influence of the initial conditions requires extensive further research. In the remainder of this section I will report the results of some experiments that give an idea of the overall tendencies though.

If the first training corpus contains no case marking at all (a somewhat unrealistic scenario, given that the **GEN** supplies case morphemes – perhaps this models the development of a language immediately after some other devices have been reanalyzed as case morphemes), the overall development is similar to the previous set up. The ranking that BiGLA induces from the initial corpus places STRUC extremely high (at 35.25), while the constraints that favor case marking are placed much lower, thus reflecting the absence of case marking. Still, the Aissen subhierarchies are respected, with *(su/a/z) at −21.33, *(su/i/z) at 4.38, *(ob/a/z) at 1.26 and *(ob/i/z) at −21.07. During the following 50 generations *STRUC is constantly lowered until it assumes a position half-way between the harmonic and the disharmonic constraints. The ranking that thus emerges is qualitatively identical to the steady state that was reached after 280 generations in the previous experiment. On the corpus side, this means that the probability of a disharmonic NP to be case marked gradually rises from 0 percent to 100 percent within 50 generations,

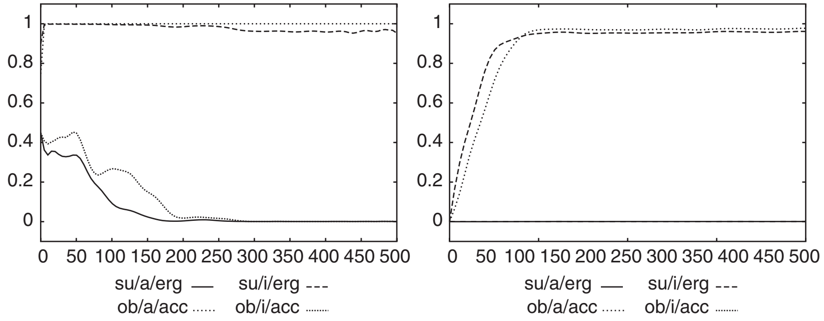


Figure 6 Probabilities of case marking

while harmonic NPs remain obligatorily unmarked. Again, the emerging split ergative system is a steady state. The change of the case marking probabilities over time is depicted in the right graphics of Figure 6.

So if the initial training corpus does not display a correlation between animacy and case marking, the iteration of bidirectional learning with the said constraint set and lexicon shows an inherent tendency towards split ergative systems.

It was mentioned in the beginning that DCM is a strong universal tendency. There are very few languages with an inverse DCM pattern. This is predicted by the assumption of Aissen's universal subhierarchies: there cannot be a language that marks animate subjects with higher probability than inanimate ones, say. It is revealing to run the BiGLA on a training corpus with such an (allegedly impossible) pattern. I did a simulation with a training corpus where all and only the harmonic NPs were case marked. The development of the constraint ranking and case marking probabilities is given in the Figures 7 and 8. The BiGLA in fact learns the inverse pattern, that is, it comes up with a grammar where the Aissen subhierarchies are reversed: *(su/a/z) $\gg$ *(su/i/z) and *(ob/i/z) $\gg$ *(ob/a/z). Accordingly, the language that is learned in the first generation marks almost all harmonic NPs but nearly no disharmonic ones. So UG admits such a language, and it is also learnable. However, it is extremely unstable. After 15 generations the Aissen subhierarchies emerge and remain stable for the remainder of the simulation (which ran over 1000 generations). Nonetheless, the case marking patterns changed dramatically after that. For about 100 generations after

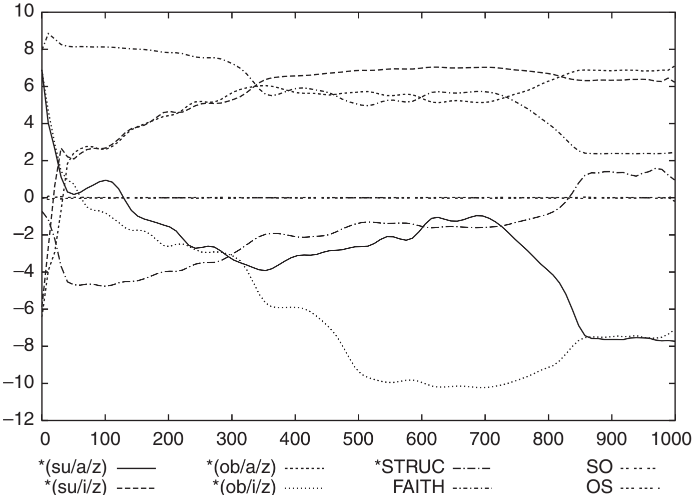


Figure 7 The future of anti-DCM: constraint rankings

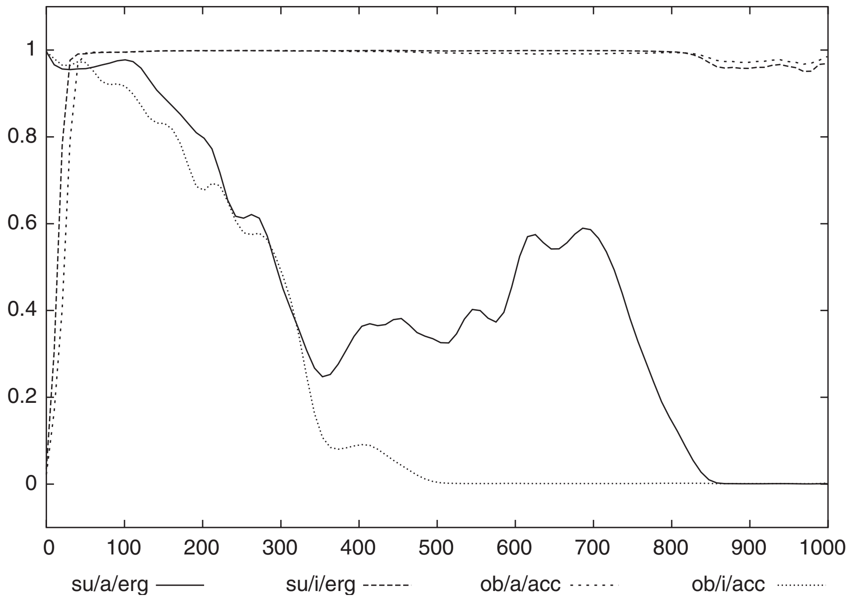


Figure 8 The future of anti-DCM: case marking probabilities

the emergence of the Aissen hierarchies, case marking is virtually obligatory for all NPs. This corresponds to a ranking were *STRUC is ranked very low. This phase is followed by a smooth raising of *STRUC, accompanied by a simultaneous lowering of *(su/a/z) and *(ob/i/z), until all three constraints are roughly at the same level. This means that case marking of harmonic NPs becomes optional while marking of disharmonic NPs remains obligatory. During the subsequent 500 generations, the symmetry between subjects and objects is broken. Accusative marking of inanimate NPs is totally lost, while ergative marking of animate NPs stays optional. After a final crisis where *(su/a/z) is lowered and hence ergative marking of animates is lost, the system also enters the steady state of split ergativity.

A large number of further simulations indicated that split ergativity is in fact the only stable state under the side conditions used here, that is, the constraint set, the generator and the relative probabilities of the possible interpretations.

While these simulations establish a connection between the statistical patterns of language use and the independently motivated constraint hierarchies postulated by Aissen, the experimental results are at odds with the actual typological tendencies. Languages with split ergativity are a minority among the languages of the world. The majority of languages follows a nominative-accusative pattern, often combined with DOM. It is a matter of dispute whether pure (morphological) ergative languages exist at all, and in any case they are very rare. How do these facts relate to the predictions of iterated learning? I will conclude this section with some speculations about the typology of case marking patterns within the paradigm of iterated learning using BiGLA.

The generator relation that was used in the above experiments represents a typologically marked language type because each NP has both an ergative and an accusative form next to the unmarked form. Such tripartite systems exist but are very rare. In most languages, each NP has at most two morphological forms for the syntactic core functions. In most split ergative languages, some NPs have a special ergative and other NPs a special accusative form next to the unmarked one, but no NP has both. So another plausible approximation to a lexicon would stipulate only two morphological forms for each NP, unmarked and marked, and leave the interpretation of the marked form as ergative or accusative to the constraint ranking.[20]

In this set-up, each transitive clause type has four morphological variants because both NPs can be either marked or unmarked each. We still have eight possible meanings. A training corpus with 50 percent probability of case marking for each NP type (using the SAMTAL distribution of meanings) thus looks as in Table 8. Here "M" stands for "marked".

In the previous set-up, the interpretation of the case morphemes was taken care of by the constraints FAITH. Since here we only have one case morpheme, this constraint has to be split into two, one favoring an accusative and one an ergative interpretation of this morpheme.

Table 8 Training corpus

	M-M	**M-Z**	**Z-M**	**Z-Z**
su/a-ob/a	1.19	1.19	1.19	1.19
su/a-ob/i	10.50	10.50	10.50	10.50
su/i-ob/a	0.07	0.07	0.07	0.07
su/i-ob/i	0.74	0.74	0.74	0.74
ob/a-su/a	1.19	1.19	1.19	1.19
ob/a-su/i	0.07	0.07	0.07	0.07
ob/i-su/a	10.50	10.50	10.50	10.50
ob/i-su/i	0.74	0.74	0.74	0.74

6.1: m ⇒ su: *Marked NPs are subjects*

6.1: m ⇒ ob: *Marked NPs are objects*

The development of the constraint rankings under this set-up is given in the first graphics of Figure 9.

Here it takes about 400 generations before a steady state is reached. The stable ranking is virtually categorical with three strata, namely:

{*(su/i/z), *(ob/a/z)} ≫ {m⇒su, m⇒ob, *STRUC, SO, OS} ≫ {*(su/a/z), *(ob/i/z)}

This ranking corresponds to a split ergative pattern. Systematic experimentation showed that as in the previous setup, split ergativity is in fact the only steady state, regardless of the nature of the initial training corpus.

However, the dynamics of the system are very sensitive to the relative frequencies of the different meanings. The emergence of Aissen's subhierarchies is due to the fact that there are much more clauses of the type "animate subject–inanimate object" than the inverse type. The clauses where both arguments are of the same animacy are irrelevant here. Their relative frequency is decisive for the precise nature of the steady states though. In the SAMTAL corpus, the number of clauses where both arguments are animate (300) has the same order of magnitude as the number of clauses with two inanimate arguments (186). If we look, for instance, at definiteness instead, this is different. Here the absolute frequencies are as in Table 9.

There are about 60 times as many clauses with two definite arguments as clauses with two indefinite NPs. Feeding a training corpus with these relative frequencies and 50 percent probability of case marking for each NP type into iterated BiGLA gives a qualitatively different trajectory than in the previous experiment. It is given in the second graphics of Figure 9.

Here the system reaches a steady state after about 70 generations. The emerging ranking is the following (where "*(ob/d/z)" stands for "Avoid unmarked definite objects!" etc.):

{*(obj/d/z), m⇒obj} ≫ *(obj/i/z) ≫ {*(subj/i/z), SO, OS} ≫
*STRUC ≫ *(su/d/z) ≫ m⇒su

Table 9 Frequencies of clause types with respect to definiteness

	subj/def	subj/indef
obj/def	1806	24
obj/indef	1292	29

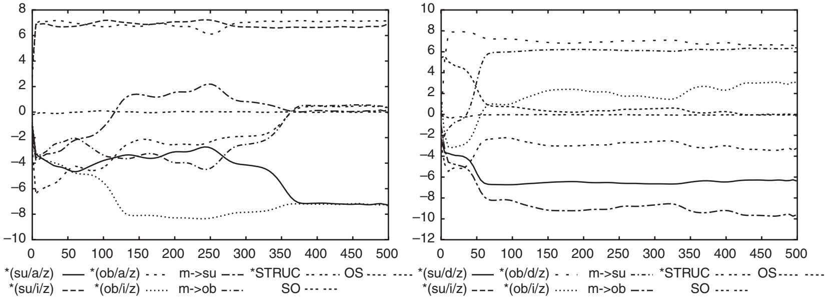


Figure 9 Simulation using only two forms per NP: animacy and definiteness

This grammar seems to describe a language with obligatory object marking and DSM. However, recall that **GEN** only supplies one case morpheme here, and the subhierarchy m⇒obj $\gg$ m⇒su ensures that this morpheme is unequivocally interpreted as accusative. Thus ergative marking is impossible and the constraint ranking describes a language with obligatory object marking and no subject marking.

To sum up the findings from this section, we may distinguish several types of case marking patterns according to their likelihood. Most unlikely are languages that violate UG, that is, where there is no constraint ranking that describes such a language. If we assume a UG as above (i.e., the **GEN** and set of constraints discussed in the previous section), there can't be a language where either both subject and object or neither are case marked. (Feeding such a corpus into BiGLA leads to a language where about 60 percent of all clauses contain exactly one case marker.) Note that it is extremely unlikely but not impossible to find a corpus with this characteristics, because this language is a subset of many UG-compatible languages. Such a corpus would be highly unrepresentative though.

The next group consists of languages that correspond to some constraint ranking, but are not learnable in the sense that exposing the BiGLA to a sample from such a language leads to a grammar of a substantially different language. A language without any case marking would fall into this category (provided **GEN** supplies case marking devices). There is a constraint ranking which describes such a language, namely:

$$\text{*STRUC} \gg \{\text{OS, SO}\} \gg \{\text{*(su/a/z), *(su/i/z), *(ob/a/z), *(ob/i/z)}\} \gg \text{FAITH}$$

However, if the BiGLA is exposed to a sample from this language, it comes up with a substantially different ranking, namely:

$$\text{*STRUC} \gg \{\text{OS, SO, FAITH}\} \gg \{\text{*(su/a/z), *(su/i/z), *(ob/a/z), *(ob/i/z)}\}$$

As can be seen from Figure 6 (p. 277), this corresponds to a language without *structural* case marking. (Structural case marking only evolves in the second generation.) However, 16.9 percent of the NPs in a sample corpus drawn from this language do carry case marking nevertheless. In other words, this language has *pragmatic* case marking.

The third group consists of languages that are both in accordance with UG and learnable (in the sense that the BiGLA reproduces a language with similar characteristics), but diachronically unstable. This means that the BiGLA acquires a language that is similar but not entirely identical to the training language, and that the deviation between training language and acquired language always goes in the same direction. Diachronically, this leads to a change of language type after some generations. This can be observed most dramatically with languages with inverse DCM (cf. Figure 8, p. 279). There the language type switches from inverse split ergativity to obligatory case marking within less than 20 generations.

There are different degrees of instability. In the third experiment reported above, a pattern with categorical DOM and optional DSM would last as long as 400 generations before it changed to categorical split ergativity.

The most likely language types are those that are diachronically stable and are additionally the target of diachronic change in many cases. The experiments conducted so far indicate that there is exactly one such steady state for each experimental set-up – split ergativity in the first two and nominative-accusative in the third scenario.[21]

Schematically expressed, this predicts the following hierarchy of language types according to their likelihood:

1. *diachronically stable and target of diachronic change*: split ergative (first two scenarios), nominative-accusative (third scenario)
2. *diachronically moderately stable*: optional DSM paired with categorical DOM (first scenario)
3. *diachronically very unstable*: inverse DCM
4. *unlearnable*: no case marking, random case marking
5. *not UG-conform*: zero or two case markers per clause

Given the extremely coarse modeling of the factors that determine case marking in our experiments and the fact that the experiments all depend on a probability distribution over meanings that is based on just one corpus study, these results have to be interpreted with extreme caution. They fit the actual patterns of typological variation fairly well though, so it seems worthwhile to pursue this line of investigation further.

11 Conclusion and open questions

In this chapter I proposed a revised version of Boersma's Gradual Learning Algorithm. It incorporates the concept of bidirectional optimization in two ways. First it uses a notion of optimality of an input–output pair that takes both the hearer perspective and the speaker perspective into account. Second, learning is thought of as bidirectional as well. The learner gradually adjusts both its production and its interpretation preferences to the observations.

The working of this bidirectional Gradual Learning Algorithm was applied to Aissen's theory of differential case marking. It could be shown that the constraint subhierarchies that Aissen simply assumes to be universal emerge automatically via learning if the training corpus contains substantially more harmonic meanings then disharmonic ones. This connection between harmony and frequency has been pointed out and used in ZJ's approach before. The present system diverges from ZJ in assuming that learning mediates between statistical biases in the language use and grammatical biases as expressed by the Aissen hierarchies, while ZJ simply identify these biases.

Several computer simulations confirmed the correlation between the statistical patterns of usage in a training corpus and the characteristics of the grammars induced from these corpora by the BiGLA.

In these experiments, only the correlation between grammatical functions and the binary contrast animate/inanimate in simple transitive active clauses was studied. Further investigations will have to use more informed models. In particular the effect of using a more articulated and perhaps two-dimensional substantive hierarchy (the combination of the definiteness hierarchy with the animacy hierarchy) as well as the effect of diathesis should be studied.

There are also several theoretical questions pertaining to the BiGLA to be addressed. The most important one is the problem of convergence of learning. By definition, a learning algorithm for a stochastic language should converge to a grammar for the learned language provided the training corpus is a representative sample of the language. The BiGLA obviously does not have this property; otherwise every language type would be stable. So is it adequate to call the BiGLA a learning algorithm to start with?

There are several points involved here. First, since the BiGLA is based on a version of bidirectional StOT, it is only supposed to learn languages that are described by a grammar from this class. That non UG-conforming languages are not learned is thus no problem. However, there are languages that correspond to some constraint ranking, but yet the BiGLA returns the grammar of a language that slightly or massively differs from the training language. The unidirectional GLA does not have this property. However, the convergence condition just requires that a learning algorithm maps *representative* samples of a language to a grammar for that language. In unidirectional StOT, this means that the different possible outputs for a given input are distributed according to the conditional probabilities that the grammar assigns to them. The relative frequencies of the inputs (= meanings) has no impact on the learning result. This is different from bidirectional learning.

Here the relative frequencies of the different meanings of a given form in the training corpus also have to converge towards their grammatically determined conditional probabilities to ensure convergence of learning. Another way to state this point is this: a StOT grammar defines a probability distribution over meaning–form pairs, and a representative sample of a language has to mirror these probabilities in frequencies. In our experimental set-up, however, the marginal probabilities of the different meanings were determined by extra-grammatical factors (the relative frequencies from SAMTAL). So the conditional probabilities of the different forms for a given meaning were matched by relative frequencies, but not the probabilities of the different meanings. Hence the BiGLA only converges towards a grammar of the training language if the SAMTAL-probabilities of meanings coincide with the probabilities assigned by the grammar. This is only the case if the least marked meanings are the most frequent ones. (This is the theoretical base for the correlation between frequencies and language types that is inherent in the BiGLA.)

Still, the grammar for the language without case marking mentioned above is in equilibrium in this sense, and yet it is not learnable by the BiGLA. How is this possible? The problem here is that during the learning process the hypothesized grammar fits the training corpus better and better, but it is not guaranteed that the difference to a real grammar becomes arbitrarily small. There are several remedies possible here, but perhaps this failure to converge with certain languages is not such a severe disadvantage after all. It should be noted that a language without case marking is extremely dysfunctional. On average only 50 percent of all utterances are interpreted correctly by the hearer. The language that the BiGLA acquires is better adapted to usage – it is due to pragmatic case marking that more than half of all utterances get their message across. So there is also a tendency towards functionality inherent in the BiGLA, and it meets the convergence condition for a stochastic learning algorithm only for languages that are functionally adapted in a certain way. The exact content of this condition is a subject for further research.

Both tendencies that are "built into" the BiGLA – frequent meanings should be unmarked meanings, and functional languages are better than dysfunctional ones – have been identified as important linguistic factors time and again by functional linguists (see, for instance, the discussions in Haspelmath, 1999, 2002). I expect that formal learning theory and functional linguistics can profit from each other a great deal, and I hope that the present chapter illustrates the fertility of such an alliance.

Acknowledgments. This chapter emerged from a series of discussions I had with Judith Aissen and Joan Bresnan. I also profited a lot from the feedback I got from the editors of this volume and two anonymous referees. Last but not least, I thank Jason Mattausch for correcting my English.

Notes

1. Here and throughout the chapter, I consider the morphological form of the subject of an intransitive clause as unmarked, and case marking that deviates from it as marked.
2. See Aissen (2000) for a more elaborate discussion, examples and references.
3. By this I mean the partial order over the Cartesian product of the domain of the two scales, where $\langle a_1, b_1\rangle \geq \langle a_2, b_2\rangle$ iff $a_1 \geq a_2$ and $b_1 \geq b_2$.
4. Here and henceforth, I use the term "subject" to refer both to the single argument of an intransitive verb and to the controller/agent argument of transitive verb. "Object" refers to the non-subject argument of a simple transitive verb. While this terminology expresses a bias towards accusative systems and against ergative systems, no real harm is done by this in the context of this chapter because it does not deal with intransitive clauses.
5. Dixon (1994, p. 90) gives two examples: the Australian language Arrernte has an inverse split ergativity system for pronouns – only first person pronouns are marked as subjects, while all other pronouns are unmarked as subjects but marked as objects. Nganasan (from the Samoyedic group of the Uralic family) has inverse DOM, i.e., full nouns but not pronouns are case marked as objects.
6. In other languages, the impact of the grammar on these quantities might be considerable. To clearly separate the usage patterns from grammatical features of the language studied, one has to look at the correlation between animacy/definiteness and *semantic* roles. This has to be left for future research.
7. The resemblance to optimal coding in the sense of information theory is striking. Shannon (1948) showed that an optimal coding must assign long codes to rare events and short codes to frequent ones.
8. The terminology I use here differs somewhat from ZJ, but is more in line with the bulk of the OT literature.
9. The notion of bidirectionality given in the definition differs from Blutner's, which treats speaker and hearer as totally symmetrical. Also, I am again deviating somewhat from the original formulation in ZJ in a way that makes no difference for their general point.
10. Boersma (1998) and Boersma and Hayes (2001) prefer values around 100 while I find values around 0 easier to work with. Since only the distance between constraint values matters and not the values as such, this makes no real difference.
11. In the graph of a normal distribution (see Figure 1), the mean corresponds to the center where the value of the function is at its maximum, and the standard deviation is the distance between the mean and the points on both sides where the shape of the curve changes from concave to convex.
12. I use C1, C2 etc. both as names of constraints and as stochastic variables over the actual values of these constraints.
13. To be precise, the dependency is the distribution function of a normal distribution with mean = 0 and standard deviation = $2\sqrt{2}$; cf. Boersma (1998), p. 284.
14. Boersma (1998) assumes that the plasticity value decreases over time. This is in fact essential to ensure that the algorithm converges. Keeping the plasticity constant lets the algorithm oscillate around the grammar to be learned without getting closer. For all practical purposes, a small constant value for plasticity is good enough though.
15. To take a simple example, suppose there are two inputs, i_1 and i_2 and one output, o. **GEN** relates both inputs to the single output. There is only one constraint that

is fulfilled by $\langle i_1, o\rangle$ but violated by $\langle i_2, o\rangle$. Hence $\langle i_1, o_{\mathrm{i}}\rangle < \langle i_2, o\rangle$, and so $\langle i_1, o\rangle$ is hearer-optimal while $\langle i_2, o\rangle$ is not. There is no hearer-optimal, and thus no optimal, output for i_2.

16. The idea to implement bidirectional optimality by using hearer-optimality as a constraint within a speaker oriented evaluation mechanism is inspired by Beaver (forthcoming). There a version of hearer-optimality is a regular constraint that can even be outranked by other constraints. I'm a bit more conservative here by treating bidirectionality as a part of the evaluation component; so it can never be outranked, and it is not subject to stochastic perturbation.
17. In Chapter 15 of Boersma (1998), Boersma also considers a purely hearer oriented version of GLA. There the learner only compares competing interpretations for the observed form. The idea of *bidirectional* learning, i.e., of simultaneous speaker-oriented and hearer-oriented learning is to my knowledge new though.
18. Due to the symmetry of the training corpus with respect to linking, OS and SO are promoted and demoted by the same amount and both remain close to 0.
19. In a corpus that was generated by the grammar of the 500th generation, more than 95 percent of all NPs follow the split ergative pattern.
20. For the purposes of this chapter, I equate the generator relation with the lexicon and hence do not assume the generator to be universal. A more refined model of learning has thus to include the acquisition of the generator as well. For the time being, I ignore this issue for the sake of simplicity.
21. It is of course possible to construct artificial scenarios with several equilibria due to perfect symmetry.

References

Adams, K., and A. Manaster-Ramer (1988). Some questions of topic/focus choice in Tagalog. *Oceanic Linguistics* 27, 79–101.

Aloni, M. (2001). Pragmatics for propositional attitudes. In R. van Rooy and M. Stokhof (Eds), *Proceedings of the Thirteenth A'dam Colloquium*. Amsterdam: ILLC.

Aissen, J. (1999). Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17, 673–711.

——(2000). Differential object marking: iconicity vs. markedness. Manuscript, UCSC.

——and J. Bresnan (2002). OT syntax and typology. Course material from the Summer School on Formal and Functional Linguistics. University of Düsseldorf.

Anderson, S. R. (2000). Towards an optimal account of second-position phenomena. In J. Dekkers, F. van der Leeuw and J. van de Weijer (Eds), *Optimality Theory: Phonology, Syntax and Acquisition* (pp. 302–33). Oxford: Oxford University Press.

Anttila, A., and V. Fong (2000). The partitive constraint in Optimality Theory. *Journal of Semantics* 17, 281–314.

Archangeli, D., and D. T. Langendoen (Eds) (1997). *Optimality Theory: An Overview*. Malden, MA and Oxford: Blackwell.

Ariel, M. (1990) *Accessing NP Antecedents*. Croom Helm Linguistics Series.

——(1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics* 16 (5), 443–64.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer Academic Publishers.

——(1999). Discourse and the focus/background distinction. In P. Bosch and R. van der Sandt (Eds), *Focus: Linguistic, Cognitive, and Computational Perspectives* (pp. 247–67). Cambridge: Cambridge University Press.

——and A. Lascarides (1998). Bridging. *Journal of Semantics* 15, 83–113.

——I. Sher and M. Williams (2001). Game theoretic foundations for Gricean constraints. Paper presented at the Thirteenth Amsterdam Colloquium, Amsterdam.

Asudeh, A. (2001). Linking, optionality and ambiguity in Marathi: an Optimality Theory analysis. In P. Sells (Ed.), *Formal and Empirical Issues in Optimality Theoretic Syntax* (pp. 257–312). Stanford, CA: CSLI Publications.

Atlas, J., and S. Levinson (1981). It-clefts, informativiteness and logical form. In P. Cole (Ed.), *Radical Pragmatics* (pp. 1–61). New York: Academic Publishers.

Axelrod, R. (1984). *The Evolution of Co-Operation*. London: Penguin.

Baković, E., and E. Keer (2001). Optionality and ineffability. In G. Legendre, J. Grimshaw and S. Vikner (Eds) (2001). *Optimality Theoretic Syntax* (pp. 97–112), Cambridge, MA: MIT Press.

Bar-Hillel, Y., and R. Carnap (1953). Semantic information. In *Proceedings of the Symposium on Applications of Communication Theory*. London: Butterworth Scientific Publications.

—— ——(1964). An outline of a theory of semantic information, In Y. Bar-Hillel (Ed.), *Language and Information* (pp. 221–74). Cambridge: Addison Wesley.

Barwise, J., and R. Cooper (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy* 4, 159–219.

Bayer, J. (1996). *Directionality and Logical Form*. Dordrecht: Kluwer Academic Publishers.

Beaver, D. I. The optimization of discourse anaphora. *Linguistics and Philosophy* (forthcoming).

——and B. Clark (2002). The proper treatment of focus sensitivity. In C. Potts and L. Mikkelson (Eds), *Proceedings of WCCFL* 21 (pp. 15–28). Somerville, MA: Cascadilla Press.

Beghelli, F., and T. Stowell (1997). Distributivity and negation. In A. Szabolcsi (Ed.), *Ways of Scope Taking* (pp. 71–107). Dordrecht: Kluwer.

Bloom, D. B. (1999). They're freezing in Russia: a typology of word order "freezing" in Russia. Manuscript, Stanford University, CA.

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics* 15, 115–62.

——(2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17, 189–216.

——(2002). Lexical semantics and pragmatics. *Linguistische Berichte* 10, 27–58.

——and G. Jäger (1999). Competition and interpretation: the German adverbs of repetition. Available at http://www2.hu-berlin.de/asg/blutner/.

——A. Leßmöllmann and R. van der Sandt (1996). Conversational implicature and lexical pragmatics. Paper presented at the AAAI Spring Symposium on Conversational Implicature, Stanford, CA.

Bobaljik, J. (2002). A-chains at the PF interface. *Natural Language and Linguistic Theory* 20, 197–267.

Boersma, P. (1998). *Functional Phonology*. The Hague: Holland Academic Graphics.

——and B. Hayes (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32, 45–86.

——and D. Weenink (2000). Praat computer program. Online, Institute of Phonetic Sciences, University of Amsterdam: www.fon.hum.uva.nl/praat.

Bolinger, D. D. (1978). Asking more than one thing at a time. In H. Hiż (Ed.), *Questions* (pp. 107–50). Dordrecht: D. Reidel.

Bosch, P. (1983). *Agreement and Anaphora: A Study of the Role of Pronouns in Syntax and Discourse*. London and New York: Academic Press.

Bossong, G. (1985). *Differentielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Günther Narr Verlag.

Bouchard, D. (1984). *On the Content of Empty Categories*. Dordrecht: Foris.

Bresnan, J. (2001a). Explaining morphosyntactic competition. In M. Baltin and C. Collins (Eds), *Handbook of Contemporary Syntactic Theory* (pp. 11–44). Oxford: Blackwell.

——(2001b). The emergence of the unmarked pronoun. In G. Legendre, J. Grimshaw and S. Vikner (Eds), *Optimality-Theoretic Syntax* (pp.113–42). Cambridge, MA: MIT Press.

——and A. Deo (2001). Grammatical constraints on variation: *'Be'* in the *survey of English dialects* and (stochastic) Optimality Theory. Manuscript, Stanford University, CA. Available online at http://www.lfg.stanford.edu/bresnan/download.html.

——S. Dingare and C. Manning (2001). Soft constraints mirror hard constraints: voice and person in English and Lummi. In M. Butt and T. H. King (Eds), *Proceedings of the LFG01 Conference* (pp. 13–32). Stanford, CA: CSLI Publications.

Büring, D. (2001). Let's phrase it: focus, word order and prosodic phrasing in German double object constructions. In G. Müller and W. Sternefeld (Eds), *Competition in Syntax* (pp. 69–106). Berlin: Mouton de Gruyter.

Burzio, L. (1991). The morphological basis of anaphora. *Journal of Linguistics* 27, 81–105.

——(1998). Anaphora and soft constraints. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis and D. Pesetsky (Eds), *Is the Best Good Enough?*. Cambridge, MA: The MIT Press.

Cable, S. (2002). Hard constraints mirror soft constraints! Bias, Stochastic Optimality Theory, and split-ergativity. Manuscript, University of Amsterdam.

Carden, G., and W. A. Stewart (1988). Binding theory, bioprogram and creolization: evidence from Haitian creole, *Journal of Pidgin and Creole Languages* 3, 1–67.

Carston, R. (2002). Linguistics meaning, communicated meaning and cognitive pragmatics. *Mind and Language* 17 (1/2), 127–48.

——(2003a). Explicature and semantics. In S. David and B. Gillon (Eds), *Semantics: A Reader*. Oxford: Oxford University Press.

——(2003b). Relevance theory and the saying/implicating distinction. In L. Horn and G. Ward (Eds), *Handbook of Pragmatics*. Oxford: Blackwell.

Choi, H.-W. (1999). *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.

Chomsky, N. (1980). On binding, *Linguistic Inquiry* 11, 1–46.

——(1981). *Lectures on Government and Binding*. Dordrecht: Foris.

——(1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Clark, H. (1987). Relevance to what? *Behavioral and Brain Sciences* 10, 714–15.

Cole, P. (1981). *Radical pragmatics* (Ed.). New York: Academic Press.

Croft, W. (2001). *Radical Construction Grammar*: *Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Dalrymple, M. E. (1993). *The Syntax of Anaphoric Binding*. Stanford, CA: CSLI Publications.

——M. Kanazawa, Y. Kim, S. Mchombo and S. Peters (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy* 21, 159–210.

Dekker, P., and R. van Rooy (2000). Bi-directional optimality theory: an application of game theory. *Journal of Semantics* 17, 217–42.

Dekkers, J., F. van der Leeuw and J. van de Weijer (2000). *Optimality Theory: Phonology, Syntax and Acquisition* (Eds). Oxford: Oxford University Press.

Diessel, H. (2000). *Demonstratives, Form, Function and Grammaticalization*. Amsterdam: John Benjamins.

Diesing, M. (1996). Semantic variables and object shift. In H. Thráinsson, S. Epstein and S. Peter (Eds), *Studies in Comparative Germanic Syntax II* (pp. 66–84). Dordrecht: Kluwer.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge: Cambridge University Press.

Ducrot, O. (1980). *Les Echelles Argumentatives*. Paris: Minuit.

——(1973). *La peuvre et le dire*. Paris: Mame.

Dugdale, N., and C. F. Lowe (2000). Testing for symmetry in the conditional discriminations of language-trained chimpanzees. *Journal of the Experimental Analysis of Behavior* 73 (1), 5–22.

Erteschik-Shir, N. (2001). P-syntactic motivation for movement. *Working Papers in Scandinavian Syntax* 68, 49–73.

Faltz, L. M. (1985). *Reflexivization: A Study in Universal Syntax*. New York and London: Garland.

Fanselow, G., and C. Féry. Ineffability in grammar. *Linguistische Berichte*, special issue: *Resolving Conflicts in Grammars: Optimality Theory in Syntax, Morphology and Phonology* (forthcoming).

Farmer, A., and M. Harnish (1987). Communicative reference with pronouns. In J. Verschueren and M. Bertuccelli-Papi (Eds), *The Pragmatic Perspective* (pp. 547–65). Amsterdam and Philadelphia, PA: John Benjamins.

Fauconnier, G. (1975). Polarity and the scale principle. In *Papers of the Eleventh Regional Meeting* (pp. 188–99). Chicago: Chicago Linguistic Society.

Fintel, K. von (1994). *Restrictions on Quantifier Domains*. Ph.D. thesis, University of Massachusetts, Amherst, MA.

Fodor, J. A., and Z. W. Pylyshyn (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71.

Fortescue, M. (1984). *West Greenlandic*. London: Croom Helm.

Fraurud, K. (2000). Demonstrativer i svensk sakprosa, In Denna – den här – den där, Om demonstrativer i tvärsprfiåklig belysning. *ASLA Information* 26 (2). Uppsala: Universitetstryckeriet.

Fry, J. (2001). *Ellipsis and wa-Marking in Japanese Conversation*. Ph.D. thesis, Stanford University, CA.

Gärdenfors, P. (1988). *Knowledge in Flux, Modeling the Dynamics of Epistemic States*. Cambridge, MA: MIT Press.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. New York: Academic Press.

——and D. Good (1982). On a notion of relevance. In N. Smith (Ed.), *Mutual Knowledge* (pp. 88–100). New York: Academic Press.

Gibson, E., and K. Broihier (1998). Optimality theory and human sentence processing. In P. Barbossa, D. Fox, P. Hagstrom, M. McGinnis and D. Pesetsky (Eds), *Is the Best Good Enough*: *Optimality and Competition in Syntax* (pp. 157–91), Cambridge, MA: The MIT Press.

Gomez-Txurruka, I. The natural language conjunction "and". *Linguistics and Philosophy* (forthcoming).

Good, L. (1950). *Probability and the Weighing of Evidence*. London: Grifin.

Green, G. (1990). Differences in development of visual and auditory-visual equivalence relations. *Journal of the Experimental Analysis of Behavior* 51, 385–92.

Green, M. (1995). Quantity, volubility, and some varieties of discourse. *Linguistics and Philosophy* 18, 83–112.

Greenberg, J. (1978). How does language acquire gender markers? In J. Greenberg, C. Ferguson and E. Moravcsik (Eds), *Universals of Human Language* (pp. 47–82). Stanford, CA: Stanford University Press.

Grice, P. (1957). Meaning. *Philosophical Review* 66, 377–88.

——(1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds), *Syntax & Semantics, Volume 3: Speech Acts* (pp. 41–58). New York: Academic Press.

——(1978). Further notes on logic and conversation. In Cole (Ed.) *Syntax & Semantics, Volume 9: Pragmatics* (pp. 113–28). New York: Academic Press.

——(1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic Inquiry* 28, 373–422.

——and V. Samek-Lodovici (1998). Optimal subjects and subject universals. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis and D. Pesetsky (Eds), *Is the Best Good Enough? Optimality and Competition in Syntax* (pp. 193–219). Cambridge, MA: MIT Press.

Groenendijk, J., and Stokhof, M. (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Grosz, B., A. Joshi and S. Weinstein (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics* (pp. 44–9). Cambridge, MA.

—— —— ——(1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* 21, 203–26.

Gundel, J., N. Hedberg and R. Zacharski (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69 (2), 274–307.

Haiman, J. (1995). Grammatical signs of the divided self. In W. Abraham, T. Givón and S. Thompson (Eds), *Discourse Grammar and Typology* (pp. 213–34). Amsterdam and Philadelphia, PA: John Benjamins.

Hale, M., and C. Reiss (1998). Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry* 29, 567–615.

Hamblin, C. L. (1973). Questions in Montague Grammar. *Foundations of Language* 10, 41–53.

Hasida, K., K. Nagao and T. Miyata (1995). A game-theoretic account of collaboration in communication. In *Proceedings of the First International Conference on Multi-Agent Systems*. San Fransisco.

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft* 18 (2), 180–205.

——(2002). Explaining the ditransitive person-role constraint: a usage-based explanation. Manuscript, MPI für evolutionäre Anthropologie, Leipzig.

Hayes, S. C., and L. J. Hayes (1989). The verbal action of the listener as the basis of rule governance. In S. C. Hayes (Ed.), *Rule Governed Behavior: Cognition, Contingencies, and Instructional Control* (pp. 153–90). New York: Plenum Press.

Heim, I. (1983). On the projection problem for presuppositions. In M. Barlow, D. Flickinger and M. Westcoat (Eds), *Second Annual Westcoast Conference on Formal Linguistics* (pp. 114–26). Stanford, CA: Stanford University.

Hendriks, P., and H. de Hoop (2001). Optimality theoretic semantics. *Linguistics and Philosophy* 24, 1–32.

Herburger, E. (2000). *What Counts: Focus and Quantification*. Cambridge, MA: MIT Press.

Himmelmann, N. Tagalog. In K. Adelaar and N. Himmelmann (Eds), *The Austronesian Languages of Asia and Madagascar*. London: Curzon Press (forthcoming).

Hirschberg (1985). *A Theory of Scalar Implicature*. Ph.D. thesis, University of Pennsylvania.

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science* 3, 67–90.

Hoeksema, J., and F. Zwarts (1991). Some remarks on focus adverbs. *Journal of Semantics* 8, 51–70.

Hoop, H. de (1995). *Only* a matter of context? In M. den Dikken and K. Hengeveld (Eds), *Linguistics in the Netherlands 1995* (pp. 113–24). Amsterdam and Philadelphia: John Benjamins.

——(2000). Optimal scrambling and interpretation. In H. Bennis, M. Everaert and E. Reuland (Eds), *Interface Strategies* (pp. 153–68). Amsterdam: KNAW.

——(2001). Making sense: the problem of unintelligibility. In GAGL 44: *Making Sense: From Lexeme to Discourse* (pp. 187–94). Department of Linguistics, University of Groningen.

——and H. de Swart (2000). Optimality theoretic semantics. *Linguistics and Philosophy* 24, 1–32.

Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, University of California.

——(1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicatures. In D. Schiffrin (Ed.), *Meaning, Form and Use in Context* (pp. 11–42). Washington DC: Georgetown University Press.

——(1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

——(1996). Exclusive company: only and the dynamics of vertical inference. *Journal of Semantics* 13, 1–40.

——(2000). From ‘if’ to ‘iff’: conditional perfection as pragmatic strengthening. *Journal of Pragmatics* 32, 289–326.

Hornstein, N. (2001). *Move*! Oxford: Blackwell.

Householder, F. W. (1971). *Linguistic Speculations*. London and New York: Cambridge University Press.

Huang, Y. (2000). *Anaphora – a Cross-Linguistic Study*. Oxford: Oxford University Press.

Hyams, N., and S. Sigurjonsdottir (1990). The development of long-distance anaphora: a cross-linguistic comparison with special reference to Icelandic. *Language Acquisition* 1, 57–93.

Hyman, L. (1984). Form and substance in language universals. In B. Butterworth, and B. Comrie and Ö. Dahl (Eds), *Explanations for Language Universals* (pp. 67–85). Berlin: Mouton.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

——(1997). *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Jacobson, R. (1962). *Selected Writings 1: Phonological Studies*. The Hague: Mouton.

Jäger, G. (2002). Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information* 11, 27–451.

——and R. Blutner (2000). Against lexical decomposition in syntax. In A. Z. Wyner (Ed.), *The Israeli Association for Theoretical Linguistics: The Proceedings of the Fifteenth Annual Conference* (pp. 113–37).

Kadmon, N. (1987). *On Unique and Non-Unique Reference and Assymetric Quantification*. Ph.D. thesis, University of Masachusetts.

Kager, R. (1999). *Optimality Theory*. Cambridge: Cambridge University Press.

Kameyama, M. (1999). Stressed and unstressed pronouns: complementary preferences. In P. Bosch and R. van der Sandt (Eds), *Focus: Linguistic, Cognitive, and Computational Perspectives* (pp. 306–21), Cambridge: Cambridge University Press.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk (Ed.), *Formal Methods in the Study of Language* (pp. 277–322). Amsterdam: Mathematisch Centrum.

——and U. Reyle (1993). *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers.

——and A. Rossdeutscher (1994). Remarks on lexical structure and drs construction. *Theoretical Linguistics* 20, 97–164.

Kaplan, D. (1979). On the logic of demonstratives. *Journal of Philosophical Logic* 8, 81–9.

Karagjosova, E. (2001). Towards a comprehensive meaning of German doch. In *Proceedings of the ESSLLI 2001 Student Session*. Helsinki: ESSLLI2001.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics* 1, 181–94.

Katz, J., and J. A. Fodor (1963). The structure of semantic theory. *Language* 39, 170–210.

Kayne, R. (1994). *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Keenan, E., and B. Comrie (1977). Noun phrase accessibility and universal grammar. *Linguistic Enquiry* 8, 63–99.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, CA: CSLI Publications.

Kempson, R. (1986). Ambiguity and the semantics-pragmatics distinction. In C. Travis (Ed.), *Meaning and Interpretation* (pp. 77–103). Oxford: Blackwell.

Kenstowicz, M., and H.-S. Sohn (1998). Accentual adaptation in north kyungsang Korean. Manuscript, MIT and Kyungpook National University. Rutgers Optimality Archive-# 345.

Kiparsky, P. (1983). Word-formation and the lexicon. In F. Ingeman (Ed.), *Proceedings of the 1982 Mid-America Linguistic Conference*.

König, E. (1991). *The Meaning of Focus Particles: A Comparative Perspective*. London and New York: Routledge.

Krifka, M. (1991). A compositional semantics for multiple focus constructions. *Linguistische Berichte*, Sonderheft 4, 17–53.

——(2002). Be brief and vague! And how Bidirectional Optimality Theory allows for verbosity and precision. In D. Restle and D. Zaefierer (Eds), *Sounds and Systems: Studies in Structure and Change. A Festschrift for Theo Vennemann* (pp. 439–58). Berlin and New York: Mouton de Gruyter.

Kripke, S. (ms.). Presupposition.

Kroeger, P. (1993). *Phrase Structure and Grammatical Relations in Tagalog*. Stanford, CA: CSLI Publications.

Kuhn, J. (2001a). *Formal and Computational Aspects of Optimality-Theoretic Syntax*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

——(2001b). Generation and parsing in optimality theoretic syntax: issues in the formalization of ot-lfg. In P. Sells (Ed.), *Formal and Empirical Issues in Optimality-Theoretic Syntax* (pp. 313–66). Stanford, CA: CSLI Publications.

Lakoff, G. (1971). Pronouns and reference. In J. McCawley (Ed.), *Syntax and Semantics,* Volume 7 (pp. 275–335), New York: Academic Press.

Lee, H. (2001a). Markedness and word order freezing. In P. Sells (Ed.), *Formal and Empirical Issues in Optimality-Theoretic Syntax* (pp. 63–128). Stanford, CA: CSLI Publications.

——(2001b). *Optimization in Argument Expression and Interpretation: A Unified Approach*. Ph.D. thesis, Stanford University, CA.

——(2002). Referential accessibility and stylistic variation in OT: a corpus study. *CLS*, 38.

——(2003). Prominence mismatch and markedness reduction in word order. *Natural Language and Linguistic Theory* 21, 617–80.

Legendre, G. (2001). Masked second-position effects and the linearization of functional features. In G. Legendre, J. Grimshaw and S. Vikner (Eds) *Optimality Theoretic Syntax* (pp. 241–77), Cambridge, MA: MIT Press.

——P. Smolensky and C. Wilson (1998). When is less more? faithfulness and minimal links in *wh*-chains. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis and D. Pesetsky (Eds), *Is the Best Good Enough? Optimality and Competition in Syntax* (pp. 249–89). Cambridge, MA: MIT Press.

Levinson, S. (1983). *Pragmatics*. Cambridge University Press.

——(1987a). Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena, *Journal of Linguistics* 23, 379–434.

——(1987b). Implicature explicated? *Behavioral and Brain Sciences* 10, 722–3.

——(1989). A review of *Relevance, Journal of Linguistics* 25, 455–72.

——(1991). Pragmatic reduction of Binding Conditions revisited, *Journal of Linguistics* 27, 107–61.

——(1995). Three levels of meaning. In F. R. Palmer (Ed.), *Grammar and Meaning: Essays in Honour of Sir John Lyons* (pp. 90–115). Cambridge: Cambridge University Press.

——(2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*. Cambridge, MA: MIT Press.

Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

—— (1979). Scorekeeping in a language game. *Journal of Philosophical Logic* 8, 339–59.

Lindblom, B. (1986). On the origin and purpose of discreteness and invariance in sound patterns. In P. J. S. and D. H. Klatt (Eds), *Invariance and Variability in Speech Processes* (pp. 493–510). Hillsdale, NJ: L. Erlbaum.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H and H theory. In W. J. Hardcastle and A. Marchal (Eds), *Speech Production and Speech Modelling* (pp. 403–39). Dordrecht: Kluwer.

Lindström, E. (2000). Some uses of demonstratives in spoken Swedish. In S. Botley and A. McEnery (Eds), *Corpus-Based and Computational Approaches to Discourse Anaphora* (pp. 107–28). Amsterdam: John Benjamins

Maclachlan, A., and M. Nakamura (1997). Case-checking and specificity in Tagalog. *The Linguistic Review* 14, 307–33.

Maes, A., and L. G. M. Noordman (1995). Demonstrative nominal anaphors: a case of non-identificational markedness. *Linguistics* 33 (2), 255–82.

Maling, J. (1984). Non-clause-bounded reflexives in Modern Icelandic. *Linguistics and Philosophy* 7, 211–41.

Mann, W. C., and S. A. Thompson (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text* 8, 243–81.

Martinet, A. (1955). *Économie des Changements Phonétiques: Traité de Phonologie Diachronique*. Berne: Éditions A. Franke.

McCarthy, J., and A. Prince (1994). The emergence of the unmarked: optimality in prosodic morphology. In M. González (Ed.), *Proceedings of NELS 24* (pp. 333–79). Amherst: GLSA, University of Massachusetts.

McCawley, J. (1978). Conversational implicature and the lexicon. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 245–59). New York: Academic Press.

Merin, A. (1999). Information, relevance, and social decisionmaking. In M. d. Rijke, L. Moss and J. Ginzburg (Ed.), *Logic, Language, and Computation,* Volume 2 (pp. 179–221). Stanford, CA: CSLI.

Mey, S. de (1991). *Only* as a determiner and as a generalized quantifier. *Journal of Semantics* 8, 91–106.

Mitchell, B. (1985). *Old English Syntex*, Vols I–II. Oxford: Clarendon Press.

Mohanan, T. (1994). *Argument Structure in Hindi*. Stanford, CA: CSLI-Publications.

Montague, R. (1970). Universal grammar. *Theoria* 36, 373–98.

Müller, G. (1999). Optionality in optimality-theoretic syntax. *Glot International* 4, 3–8.

O'Connor, S. (1983). Two kinds of bound anaphora in Northern Pomo: are they logophoric? In J. E. Redden (Ed.), *Papers from the 1983, 1984, and 1985 Hokan-Penutian Languages Conferences*.

Orgun, C. O., and R. L. Sprouse (1999). From MPARSE to CONTROL: deriving ungrammaticality. *Phonology* 16, 191–224.

Paolillo, J. (2002). *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford, CA: CSLI Lecture Notes.

Parikh, P. (2000). Communication, meaning, and interpretation. *Linguistics and Philosophy* 23, 185–212.

Partee, B. H. (1991). Topic, focus and quantification. In S. Moore and A. Wyner (Eds), *Proceedings of SALT 1* (pp. 159–87). Ithaca, NY: Cornell University.

——(1999). Focus, quantification, and semantics-pragmatics issues. In P. Bosch and R. van der Sandt (Eds), *Focus: Linguistic, Cognitive, and Computational Perspectives* (pp. 213–31). Cambridge: Cambridge University Press.

Pesetsky, D. (1987). *Wh*-in-situ: movement and unselective binding. In E. Reuland and A. ter Meulen (Eds), *The Representation of (In)definiteness* (pp. 98–129). Cambridge, MA: MIT Press.

——(1997). Optimality theory and syntax: movement and pronunciation. In D. Archangeli and D. T. Langendoen (Eds), *Optimality Theory: An Overview* (pp. 134–70). Malden, MA, and Oxford: Blackwell.

——(1998). Some optimality principles of sentence pronunciation. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis and D. Pesetsky (Eds), *Is the Best Good Enough? Optimality and Competition in Syntax* (pp. 337–84). Cambridge, MA: MIT Press.

Poesio, M., and N. Modjeska (2002). The THIS-NPs hypothesis: a corpus-based investigation. Paper presented at DAARC; 4th Discourse Anaphora and Anaphor Resolution Colloquium, Lisbon, September 18–20, 2002.

Pollard, C., and I. Sag (1992). Anaphors in English and the scope of binding theory. *Linguistic Inquiry*, 23, 261–303.

Popper, K. (1959). *The Logic of Scientific Discovery*. London Hutchinson.

Poser, W. (1992). Blocking of phrasal constructions by lexical items. In I. Sag and A. Szabolcsi (Eds), *Lexical Matters* (pp. 111–30). Stanford, CA: CSLI Publications.

Postal, P. (1972). Some further limitations of interpretive theories of anaphora. *Linguistic Inquiry* 3.

Prince, E. F. (1981). Toward a taxonomy of given–new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223–55). New York: Academic Press.

Prince, A., and P. Smolensky (1997). Optimality: from neural networks to universal grammar. *Science* 275, 1604–10.

——and P. Smolensky (1993/2002). Optimality theory: constraint interaction in generative grammar. Technical report, Rutgers University and University of Colorado at Boulder, Johns Hopkins University (2nd edition only). Manuscript. Republished with minor changes at the Rutgers Optimality Archive, ROA-#537.

Reinhart, T., and E. Reuland (1991). Anaphors and logophors: an argument perspective. In J. Koster and E. Reuland (Eds), *Long-Distance Anaphora* (pp. 283–321). Cambridge: Cambridge University Press.

——(1993). Reflexivity. *Linguistic Inquiry* 24, 657–720.

Reuland, E. (2001). Primitives of binding. *Linguistic Inquiry* 32 (3), 439–92.

Rooth, M. (1985). *Association with Focus*. Ph.D. thesis, University of Massachusetts, Amherst, MA.

——(1992). A Theory of Focus Interpretation. *Natural Language Semantics* 1, 75–116.

Samek-Lodovici, V. (2002). Prosody-syntax interaction in the expression of focus. Rutgers Optimality Archive, ROA-524.

Sankofi, D and D. Rand (1999).
http://www.crm.umontreal.ca/šankofi/GoldVarb Eng.html.

Savage, L. (1954). *Foundations of Statistics*. New York: Wiley.

Schachter, P. (1993). Tagalog. In J. Jacobs, A. V. Stechow, W. Sternefeld and T. Vennemann (Eds), *Syntax. An International Handbook of Contemporary Research* (Vol. 1, pp. 1418–30). Berlin: de Gruyter.

——and F. Otanes (1972). *Tagalog Reference Grammar*. London and Berkeley, CA: University of California Press.

Schmid, T., and R. Vogel (submitted). Dialectal variation in German 3-verb-clusters: a surface-oriented OT account. Unpublished manuscript, University of Stuttgart and University of Potsdam.

Schösler, H. (2002). *Aber, aber*. M.Sc. dissertation, Computational Linguistics, Amsterdam University.

Schulz, K. (2001). *Relevanz und 'Quantity' Implikaturen*. M.Sc. dissertation, University of Stuttgart.

Schwarzschild, R. (1997). Why some foci must associate. Unpublished manuscript, Rutgers University, New Brunswick, NJ.

——(1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics* 7, 141–77.

Shannon, C. (1948). A mathematical theory of communication. *Bell Sys. Tech. Journal* 27, 379–432, 623–56.

Sells, P. (2001). *Formal and Empirical Issues in Optimality Theoretic Syntax*. Stanford, CA: CSLI Publications.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Ed.), *Grammatical Categories in Australian Languages* (pp. 112–71). Canberra: Australian Institute of Aboriginal Studies.

Smolensky, P. (1995). On the internal structure of the constraint component Con of UG. ROA 86. Handout of talk given at UCLA.

——(1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27, 720–31.

——(1998). Why syntax is different (but not really): ineffability, violability and recoverability in syntax and phonology. Handout of the talk given at the Stanford/CSLI Workshop: Is Syntax Different? Common Cognitive Structures for Syntax and Phonology in Optimality Theory. Stanford University, CA, December 12–13, 1998.

Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry* 13, 483–545.

Sperber, D., and D. Wilson (1986/1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic* 2, 447–57.

——(1978). Assertion. In P. Cole (Ed.), *Syntax and Semantics, Volume 9: Pragmatics* (pp. 315–32). New York: Academic Press.

——(1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stechow, A. von (1991). Current issues in the theory of focus. In A. von Stechow and D. Wunderlich (Eds), *Semantik/Semantics: An International Handbook of Contemporary Research* (pp. 804–25). Berlin: Mouton de Gruyter.

Stirling, L. (1993). *Switch-Reference and Discourse Representation*. Cambridge: Cambridge University Press.

Sweet, H. (1882/1953). *Sweet's Anglo-Saxon Primer*, N. Davis (Ed.). Oxford: Oxford University Press.

Tesar, B., and P. Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

Truckenbrodt, H. (1999). On the relation between syntactic phrases and phonological phrases. *Linguistic Inquiry* 30, 219–55.

Umbach, C. (2001). Contrast and contrastive topic. In I. Kruijfi-Korbayova and M. Steedman (Eds), *Proceedings of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*. ESSLLI2001.

Vallduví, E. (1992). *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Van der Does, J., and H. de Hoop (1998). Type-shifting and scrambled definites. *Journal of Semantics* 15, 393–416.

Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics* 9, 333–77.

——and B. Geurts (2001). Too. In R. van Rooy (Ed.), *Proceedings of the 13th Amsterdam Colloquium*. ILLC.

Van Kuppevelt, J. (1996). Inferring from topics: scalar implicature as topic-dependent inferences. *Linguistics and Philosophy* 19, 555–98.

Van Rooy, R. (1999). Questioning to resolve decision problems. In P. Dekker (Ed.), *Proceedings of the 12th Amsterdam Colloquium*. Amsterdam.

——(2001). Conversational implicatures. In J. van Kuppevelt and R. Smith (Eds), *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg.

——(2002). Utility, informativity, and protocols. In G. Bonanno (Ed.), *Proceedings of LOFT 5: Logic and the Foundations of the Theory of Games and Decisions*. Torino.

—— Signaling games select horn strategies. *Linguistics and Philosophy* (forthcoming).

——and K. Schulz (2003). Exhaustification. In H. Bunt (Ed.), *Proceedings of the Fifth International Workshop on Computational Semantic*. Tilburg.

Venditti, J. J., M. Stone, P. Nanda and P. Tepper (2002). Discourse constraints on the interpretation of nuclear-accented pronouns. Conference on Speech Prosody 2002.

Vikner, S. (1997). The interpretation of object shift, optimality theory, and minimalism. *Working Papers in Scandinavian Syntax* 60, 1–24.

——(2001). The interpretation of object shift and optimality theory. In G. Müller and W. Sternefeld (Eds), *Competition in Syntax* (pp. 321–40). Berlin: Mouton de Gruyter.

Visser, F. T. (1963) *An Historical Syntax of the English Language*, Part I. Leiden: E. J. Brill.

Vogel, R. (2001). Case conflict in German free relative constructions: an optimality theoretic treatment. In G. Müller and W. Sternefeld (Eds), *Competition in Syntax* (pp. 341–75). Berlin: Mouton de Gruyter.

——(2002a). Feedback optimisation: economy and markedness in OT syntax, pt. I. Unpublished manuscript, University of Potsdam.

——(2002b). Free relative constructions in OT syntax. In G. Fanselow and C. Féery (Eds), *Sonderheft Optimality Theory, Linguistische Berichte* (pp. 119–62). Hamburg: Helmut Buske Verlag.

Wasow, T., A. Perfors and D. I. Beaver. The puzzle of ambiguity. In P. Sells (Ed.), *Essays in Honor of Steve Lapointe*. Stanford, CA: CSLI Publications (forthcoming).

Webber, B. (1991). Structure and ostention and the interpretation of discourse. *Deixis, Language and Cognitive Processes* 6, 107–35.

Williams, E. (1997). Blocking and anaphora. *Linguistic Inquiry* 28, 577–628.

Wilson, C. (2001). Bidirectional opimization and the theory of anaphora. In G. Legendre, J. Grimshaw and S. Vikner (Eds), *Optimality Theoretic Syntax* (pp. 465–508). Cambridge, MA: MIT Press.

Wurzel, W. U. (1998). On markedness. *Theoretical Linguistics* 24, 53–71.

Zeevat, H. (1992). Presupposition and accommodation in update semantics. *Journal of Semantics* 9, 379–412.

——(1994). Questions and exhaustivity in update semantics. In H. Bunt (Ed.), *Proceedings of the International Workshop on Computational Semantics*. Tilburg.

——(1999). Explaining presupposition triggers. In P. Dekker (Ed.), *Proceedings of the 12th Amsterdam Colloquium*. Amsterdam.

——(2000). The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics* 17, 243–62.

——(2002). Explaining presupposition triggers. In K. van Deemter and R. Kibble (Eds), *Information Sharing* (pp. 61–87). Stanford, CA: CSLI Publications.

——and G. Jäger (2002). A reinterpretation of syntactic alignment. In D. de Jongh, M. Nilsenova and H. Zeevat (Eds), *Proceedings of the Fourth International Tbilisi Symposium on Language, Logic and Computation*. University of Amsterdam.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

Author Index

Subject Index