

ZASPiL Nr. 38 – January 2005

**On the Optimization and  
Grammaticalization of Anaphora**

Jason Mattausch

On the Optimization and Grammaticalization  
of Anaphora

Jason Mattausch

# Contents

<b>1</b>	<b>Introduction</b>	<b>6</b>
<b>2</b>	<b>Binding in Generative Grammar</b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	Syntactic Approaches to Binding Phenomena . . . . .	9
2.2.1	R-expressions . . . . .	13
2.2.2	Pronouns . . . . .	16
2.2.3	Anaphors . . . . .	18
2.3	Semantic Approaches to Binding Phenomena . . . . .	27
2.4	Summary . . . . .	32
<b>3</b>	<b>Pragmatic Approaches to Binding Phenomena</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Gricean Pragmatics . . . . .	35
3.3	Radical Pragmatics . . . . .	37
3.4	Radical Pragmatics and Anaphora . . . . .	43
3.5	Summary . . . . .	50
<b>4</b>	<b>Optimality, Superoptimality &amp; Anaphora</b>	<b>52</b>
4.1	Introduction . . . . .	52
4.2	Bidirectional OT . . . . .	54
4.3	A Pragmatic/OT Approach to Binding Phenomena . . . . .	59
4.4	Some Applications . . . . .	61
4.4.1	Introduction . . . . .	61
4.4.2	Old English . . . . .	61
4.4.3	Pidgins and Creoles . . . . .	63
4.4.4	Australian and Austronesian languages . . . . .	64
4.5	Summary . . . . .	68

<b>5</b>	<b>Bias, Stochastic OT, the GLA, &amp; Grammaticalization</b>	<b>69</b>
5.1	Introduction . . . . .	69
5.2	Optimization and Bias . . . . .	74
5.3	Stochastic OT . . . . .	81
5.4	The GLA and Grammaticalization . . . . .	84
5.5	BiGLA . . . . .	90
<b>6</b>	<b>Bias, Bidirectionality, &amp; Binding Phenomena</b>	<b>103</b>
6.1	Introduction . . . . .	103
6.2	The Basic Story . . . . .	104
6.3	The Next Generation . . . . .	119
6.4	Pattern Generalization . . . . .	133
6.5	Languages with multiple reflexivizing strategies . . . . .	141
6.6	Loose Ends . . . . .	151
	6.6.1 C-command & the Thematic Hierarchy Condition . . .	151
	6.6.2 LDAs . . . . .	154
	6.6.3 Principle C effects . . . . .	160
<b>7</b>	<b>Conclusion</b>	<b>163</b>
	General Index . . . . .	179
	Index of Authors . . . . .	181
	Index of Languages . . . . .	184

# Preface

The overall aim of this dissertation is to defend the idea that the empirical responsibilities of binding theory can be managed in a more psychologically and historically realistic way when assigned to the field of pragmatics.

I am not the first to defend such an idea, for it has already been advocated quite vigorously by some in the ‘radical pragmatics’ camp, especially Stephen Levinson and Yan Huang.<sup>1</sup> I believe though, that a pragmatic account of the behavior of bound pronouns and reflexives can be much more accurately defended if, in doing so, one makes reference to a formal theory of grammar and language learning.

In this work, I will argue that the Optimality Theory (OT) of Alan Prince and Paul Smolensky and a host of other advances in that field of research – including but not limited to Reinhard Blutner’s bidirectional OT and Gerhard Jäger’s Bidirectional Gradual Learning Algorithm – are formal tools which are excitingly well suited for such a task. In particular, I will try to show that the phenomenon that Larry Horn called the ‘division of pragmatic labor’ – whereby relatively marked forms gravitate toward relatively marked meanings – can be described as an evolutionary strategy predicted for by Jäger’s model of bidirectional learning. This in turn can give both support and clarification to the claims of Levinson and others that patterns of reflexive marking are (at least in part) manifestations of that phenomenon.

In addition to the usual introductory and conclusive sections, the dissertation consists of five chapters.

Chapter 2 discusses binding phenomena and the traditional treatments of those phenomena in the generative grammar tradition.

The third chapter is a discussion of Levinson’s ‘neo-Gricean’ work on anaphora, which suggests a pragmatic explanation for binding phenomena and the historical behavior of those phenomena based on his theory of gen-

---

<sup>1</sup>I will forgo all citations in the preface, for brevity. They can be found in the text.

eralized conversational implicatures.

In the fourth chapter, I introduce the Optimality Theory framework and Blutner's bidirectional version of it, as well as his idea that bidirectional OT can be used to recast aspects of Levinson's theory of generalized conversational implicatures in a way that both simplifies the theory and relates it to a formal theory of language comprehension and production. With this idea in mind, I will briefly sketch an OT-based picture that can mimic the empirical coverage of Levinson's pragmatic treatment of basic of binding patterns.

Chapter 5 discusses the suggestions of Henk Zeevat and Gerhard Jäger to the effect that statistical asymmetries in language use can influence grammatical knowledge and that this influence, when considered in the context of a bidirectional OT framework, can function as a vehicle for grammaticalization. In addition, I introduce Boersma's stochastic OT and his Gradual Learning Algorithm, along with Jäger's idea that both of the latter can be 'bidirectionalized' to give a formal theory of bidirectional learning that can clarify and improve upon the OT-based grammaticalization pictures advocated by Zeevat and by Seth Cable.

Prior to conclusion, the penultimate chapter applies the ideas discussed in Chapters 4 and 5 to basic binding patterns and shows how the neo-Gricean account of those patterns proposed by Levinson can be given corroboration, as well as a great deal of elucidation, when recast in the formal framework mentioned above.

The research for this dissertation was carried out primarily at the 'ZAS' – the *Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung* – in Berlin, in participation with the *Bidirektionale Optimalitätstheorie* project, done under the sponsorship of the *Deutsche Forschungsgemeinschaft*. My first thanks go to the DFG and to the members of our review committee for allowing this project to go forward.

In addition I would like to extend my thanks to the following people.

Thanks to Reinhard Blutner who served as my advisor, and to Henk Zeevat, who was like an unofficial advisor in many ways, even though he never volunteered for that job. Both of these two men have been more kind and generous to me than I can ever thank them for, so I thank them here for their helpful comments on earlier versions of this dissertation and for watching out for me in these last years.

Thanks to Manfred Krifka, who not only served as a second advisor and served as co-leader the ZAS project, but who also, in 2001-02, put me under his employ as a *Dozent* at Humboldt University and as an assistant, a position

without which I cannot imagine how I would have gotten by.

Thanks to Gerhard Jäger, the other co-leader of the ZAS project who was always willing to answer my stray questions about various things when few, if any, others would have been able to.

On a more personal note, thanks to my mother, Susan Eldred, for being so supportive all these years and making it possible for me to begin and continue studying in the first place.

Finally, thanks to Jim Levey, who was not only my dearest friend for many years, but who also, sometime in the spring of 1993, invited me to see a documentary film playing at the Nuart Theater in West Los Angeles entitled *Manufacturing Consent*, whose focus was an American political dissident named Noam Chomsky. It was some time around then that I decided to find out what a linguist was and to take Jim's advice about pursuing university studies. I dedicate this dissertation to him, because he changed my life.

Again, my deepest thanks to all of the above, and to all my other colleagues, students, and friends in Berlin, at Humboldt University, and at the ZAS, for allowing me the opportunity to work, study, and live among people with such a deep love and respect for the sciences, for humanity, and for peace.

Enjoy.

J.M.

Berlin, Germany

25 December, 2003

# Chapter 1

## Introduction

The purpose of this dissertation is to defend the idea that the empirical responsibilities of binding theory can be handled in a more psychologically and historically realistic way when assigned to the field of pragmatics. In particular, I wish to show that Optimality Theory (OT) (Prince & Smolensky, 1993), the stochastic OT and Gradual Learning Algorithm of Boersma (1998), the Recoverability OT of Wilson (2001) and Buchwald et al. (2002), and the bidirectional OT of Blutner (2000b) and Bidirectional Gradual Learning Algorithm of Jäger (2003a) can all participate in a formal framework in which one can formally spell out and justify the idea that the distributional behavior of bound pronouns and reflexives is a pragmatic phenomenon.

Recent work in the field of pragmatics, especially the ‘neo-Gricean’ work of Levinson (1991, 2000, et al.) and Huang (1994, 2000, et al.), has posed direct challenges to the ideas of those working in the generative grammar tradition, who have long treated anaphoric binding phenomena as a question of configurational relationships (in Chomsky’s case (Chomsky, 1986, e.g.)) or of semantic ones (à la Reinhart & Reuland (1993, et al.)).

It is fair to say, I believe, that many if not most (well spelled out) suggestions to the effect that binding phenomena can be explained in terms of pragmatics hinge on one crucial idea. The idea is that relatively ‘marked’ forms – i.e., those which are for some reason structurally dispreferred – tend to be used to describe or represent ‘marked’ situations – i.e., ones that are rare, unusual, or go against some contextual grain. This general idea has been invoked in some form or another by, among others, Shannon (1948), Atlas & Levinson (1981), Horn (1984), and Blutner (2000b), all of whom have suggested some framework in which a marked-form-for-marked-meaning strategy



is recognized as ‘optimal’. Such a strategy makes intuitive sense, of course, for if we assume that there is a cost, i.e., an expenditure of energy, associated with linguistic expressions then a marked-form-for-marked-meaning strategy would necessarily reduce the cost of linguistic communication in general.

Precisely how such a strategy could be relevant to anaphoric binding behavior is something I will turn to in subsequent chapters, but the basic idea of describing binding phenomena as a manifestation of that strategy rests on the ideas that (a) expressions which have grammaticalized into reflexive anaphora tend to be relatively, structurally marked forms compared to pronouns and (b) reflexive predicates are marked because, insofar as language use is concerned, they are rarer and/or more unusual than non-reflexive ones.

According to Levinson (2000, et al.), a pragmatic theory of conversational implicatures can explain why a structurally marked form and a ‘marked’ meaning would pair together in a synchronic, pragmatic sense, but his explanation of how they pair together a diachronic, evolutionary sense depends largely on the idea that such implicatures undergo a process of ‘freezing’ or ‘fossilization’ whereby the content of an implicature ‘turns into’ genuine semantic information or into a bona fide grammatical rule. One bed of pragmatic fossils, Levinson argues, has been the object of study in the generative grammar tradition, namely, binding phenomena. The crucial fossilization process, however, has never been spelled out formally.

A formal ‘theory of fossilization’ would, it seems, need to make reference to some formal theory of grammar and language learning. In recent work, Zeevat & Jäger (2002), Zeevat (2002), Cable (2002), and Jäger (2003a) have done exactly this, and all have suggested a model of grammaticalization whereby pragmatic factors can explain certain grammaticalized marking strategies. My aim is to follow these suggestions and to demonstrate a way in which the claim that basic patterns of anaphoric binding are manifestations of a marked-forms-for-marked-meanings pragmatic strategy can be stated within a formal framework. In particular, I will try to show that all of the advances in OT research mentioned above can help answer questions about the grammaticalization of anaphora and the general trends that that grammaticalization has been shown to follow.

Ultimately, I will argue that universal trends in binding phenomena and the marked-form-for-marked-meaning pattern in general can be viewed as a direct consequence of three things: universal markedness and faithfulness constraints, familiar to all OT analyses; interpretational bias – of the kind proposed by Zeevat (2002) et al.; and bidirectional learning.

The organization of the dissertation is as follows.

Chapter 2 introduces binding phenomena and surveys two alternative treatments of the phenomena within the traditional generative grammar framework, in particular, Chomsky's work on Binding Theory beginning with Chomsky (1980) and Reinhart & Reuland's alternative approach first suggested in (Reinhart & Reuland, 1991).

Chapter 3 discusses Levinson's neo-Gricean theory of generalized conversational implicatures and his pragmatic account of anaphora based on that theory.

In the fourth chapter, I introduce the Optimality Theory framework and Blutner's bidirectional version of it, as well as his idea that bidirectional OT can be used to recast aspects of Levinson's theory of generalized conversational implicatures in a way that both simplifies the theory and relates it to a formal theory of language comprehension and production. With this idea in mind, I briefly sketch an OT-based picture that can mimic the empirical coverage of Levinson's pragmatic treatment of binding phenomena.

Chapter 5 discusses the suggestions of Zeevat & Jäger (2002) and Zeevat (2002) to the effect that statistical asymmetries in language use can influence grammatical knowledge and that this influence, when considered in the context of a bidirectional OT framework, can function as a vehicle for grammaticalization. In addition, the stochastic OT and Gradual Learning Algorithm of Boersma (1998) are introduced, along with Jäger's (2003a) idea that these can both be 'bidirectionalized' to give a formal theory of bidirectional learning and clarify the OT-based grammaticalization pictures advocated by Zeevat and by Cable (2002).

Chapter 6 attempts to show how these ideas can be applied to questions surrounding the synchrony and diachrony of binding phenomena and how the neo-Gricean account of binding phenomena proposed by Levinson can be corroborated somewhat, yet clarified a great deal, when recast in the formal framework mentioned above.

Chapter 7 concludes.

# Chapter 2

## Binding Phenomena in the Generative Grammar Tradition

### 2.1 Introduction

This chapter discusses the treatment of binding phenomena in the generative grammar framework. I will not give an exhaustive summary of generative approaches to binding but will rather concentrate on two very different approaches, both of which are representative of work in that field and which differ a bit in their empirical coverage. The first is the geometrical theory of binding advocated within the Government & Binding (GB)/Principles & Parameters (P&P) framework of (Chomsky, 1980, et al.), perhaps the most well known approach to binding of all, and the second is the semantically grounded binding theory of Reinhart & Reuland (1991, et al.), which is likely the most oft-cited alternative.<sup>1</sup>

### 2.2 Syntactic Approaches to Binding Phenomena

The classical Chomskyan approach to binding phenomena is formulated in his *Binding Theory* (BT) (Chomsky, 1980, 1981, 1982, 1986). BT is meant

---

<sup>1</sup>I will forgo any discussion of the formulations of binding theory in non-GB syntactic frameworks such as LFG and HPSG, though for a comparative discussion cf. Everaert (2001), who largely downplays the differences.

to be a module of grammar that places specific configurational constraints on various types of NPs and, in doing so, regulates the referential properties of those NPs. In particular, BT is responsible for the regulation of relations between NPs in bound argument positions, (i.e., ‘*A-positions*’)<sup>2</sup> and their binders. In this sense BT is a theory of *A-binding*, where the relevant definitions are as below.<sup>3</sup>

(2.1) *(A-)Binding*

- $\alpha$  *(A-)binds*  $\beta$  iff
- a.  $\alpha$  is in an A-position
  - b.  $\alpha$  c-commands  $\beta$
  - c.  $\alpha$  and  $\beta$  are coindexed.

(2.2) *C-command*

- $\alpha$  *c-commands*  $\beta$  iff
- a.  $\alpha$  does not dominate  $\beta$
  - b.  $\beta$  does not dominate  $\alpha$
  - c. the first branching node dominating  $\alpha$  also dominates  $\beta$ .

The BT is traditionally formulated as three principles – A, B and C – each which addresses the distributional restrictions on a different type of NP, namely, ‘Anaphors’ (i.e., reflexives and reciprocals), pronouns, and full referential NPs (or ‘R-expressions’), respectively. Chomsky (1982, 78-89) argues that the notions ‘Anaphor’, ‘pronoun’, and ‘R-expression’ are not syntactic primitives, but can rather be characterized in terms of two primitive, binary features  $\pm$ Anaphoric and  $\pm$ Pronominal, where the distribution of features for overt NPs is as below.<sup>4</sup>

---

<sup>2</sup>Importantly, this excludes NPs in so-called ‘A’ positions’, i.e., those which are not associated with a  $\theta$ -role, such as in sentence like *John, everyone thinks he is a fool*, or non-argument positions as in *John met the king himself*.

<sup>3</sup>The formulation of BT presented below is most closely based on Chomsky (1981). Many revised versions of BT have been proposed in the literature on various grounds, cf. e.g., Chomsky (1986) as well as the reductionist approaches of Manzini (1983), Bouchard (1984), and Burzio (1989), inter alia, or the expansionist revisions of Lasnik (1989) and later Thráinsson (1991). It is beyond the scope of this work to introduce or debate the details of these issues, though.

<sup>4</sup>I forgo any discussion of non-overt NPs postulated in the GB or other generative grammar frameworks and will largely forgo consideration of the issues surrounding them

(2.3)		<i>R-expressions</i>	<i>pronouns</i>	<i>Anaphors</i>
	<i>Pronominal</i>	–	+	–
	<i>Anaphoric</i>	–	–	+

The BT can then be formulated in terms of the feature specifications for the various NP-types.

(2.4) *Binding Principles*

*Principle A:* An NP with the feature +Anaphoric must be bound in its governing category.

*Principle B:* An NP with the feature +Pronominal must be free in its governing category.

*Principle C:* An NP with the features –Anaphoric and –Pronominal must be free everywhere.

Principles A through C depend crucially on the notion of a local domain or ‘governing category’, which is in turn parasitic on the notions of ‘government’ and ‘accessible subject/SUBJECT’, defined below.

(2.5) *Governing Category* (GC)

The *governing category* of  $\beta$  is the minimal domain containing

- a.  $\beta$
- b. a governor of  $\beta$
- c. an accessible subject/SUBJECT for  $\beta$ .

(2.6) *Government*

$\alpha$  *governs*  $\beta$  iff

- a.  $\alpha$  does not dominate  $\beta$
- b. the lowest maximal projection that dominates  $\alpha$  also dominates  $\beta$
- c. there is no maximal projection (of a lexical head) between  $\alpha$  and  $\beta$ .

(2.7) *subject/SUBJECT*

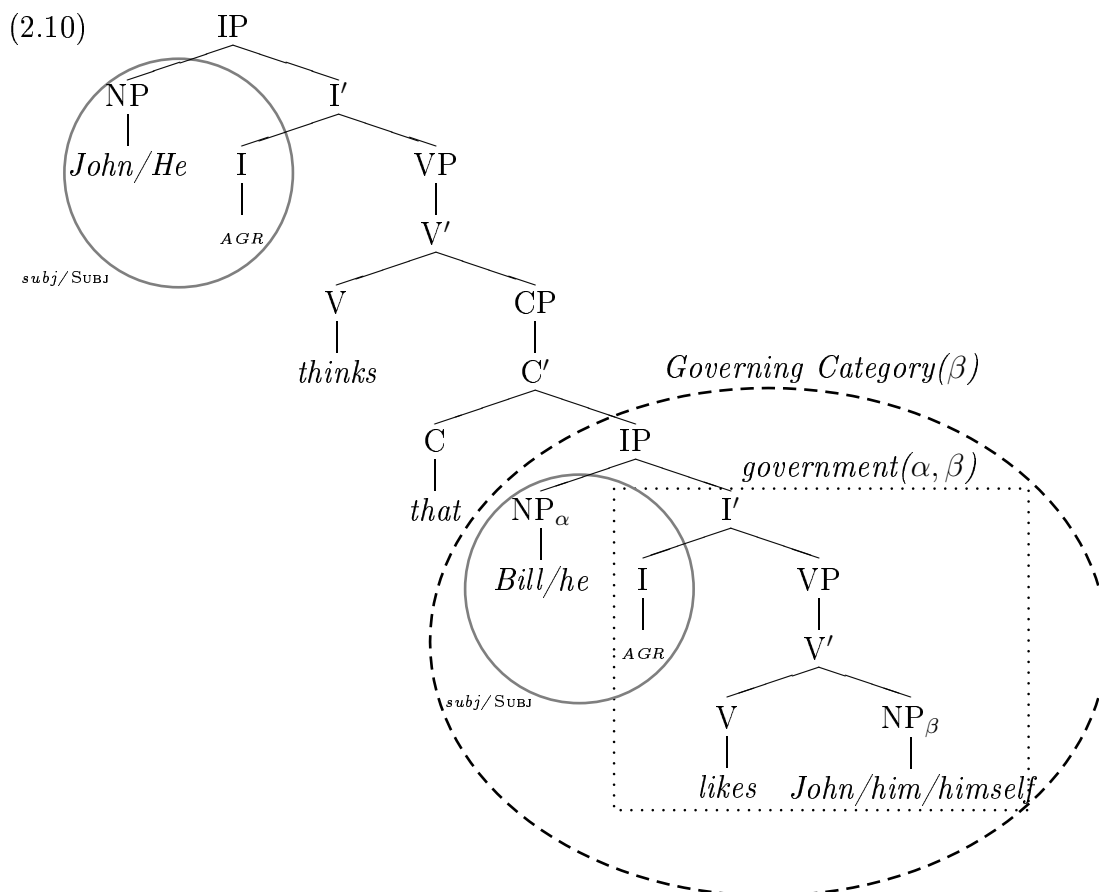
- a. subject: NP in [Spec, XP]
- b. SUBJECT corresponds to finite AGR.

---

in subsequent chapters as well, though for a critical discussion of null anaphora in the generative grammar framework, cf. Huang (2000, 50-90), who also suggests detailed pragmatic alternatives to these analyses in several works (Huang, 1994, 2000, et al.).

(2.8) *Accessible subject*/SUBJECT

$\alpha$  is an accessible subject/SUBJECT for  $\beta$  if the coindexation of  $\alpha$  and  $\beta$  does not violate any grammatical principles.<sup>5</sup>



Below some of the successes and failures of each of the three Binding Principles are individually sketched.

<sup>5</sup>As for which grammatical principles there are to be violated, one example is Chomsky's *i-within-i filter*.

(2.9) *i-within-i filter* (Chomsky, 1981, 211-212)

\*[ $A_i \dots B_i \dots$ ]

Such a filter rules out sentences which manifest circularity in reference, wherein an NP is coreferential with the NP in which it is contained, e.g., \**A picture<sub>i</sub> of itself<sub>i</sub> is on the wall.*

## 2.2.1 R-expressions

R(eferential)-expressions – such as names, (in)definite descriptions, quantifier phrases, and so on – are so-called exactly because they are viewed as being inherently referential, where “inherently referential” is typically taken to mean that they select a referent from the universe of discourse, indicating that there is some entity which is identifiable by the NP rather than merely a linguistic antecedent. On such a view, antecedents – especially intrasentential antecedents – are something which R-expressions typically resist. Consider:

- (2.11) a. John<sub>i</sub> likes John/the man<sub>\*i/j</sub>.  
b. He<sub>i</sub> likes John/the man<sub>\*i/j</sub>.  
c. John<sub>i</sub> thinks John/the man<sub>\*i/j</sub> is smart.  
d. He<sub>i</sub> thinks John/the man<sub>\*i/j</sub> is smart.

The pattern illustrated above follows directly from Principle C. However, examples have been pointed out that suggest three things: Firstly, the traditional formulation of Principle C is too strong and must permit for exceptions.

- (2.12) Classical Greek (Luke 20:25, NA26)

Ho de eipen pros autous, Toinun apodote ta Kaisaros Kaisari kai ta tou Theou to Theo.

‘And he said unto them: Render unto Caesar what is Caesar’s, and unto God what is God’s.’

- (2.13) Only Bush would vote for Bush.

Secondly, the strength of whatever Principle C-like force is responsible for the ungrammaticality of the coindexations in examples like (2.11) varies cross-linguistically. To illustrate this last point, Lasnik (1989) cites the following examples which, while structurally analogous to the English example in (2.11c), are claimed to be grammatical in Thai and Vietnamese.

- (2.14) Thai (*Ibid.*)

Coon<sub>i</sub> khit waa Coon<sub>i</sub> chalaat.

John think that John smart

‘John thinks that John is smart.’

(2.15) Vietnamese (*Ibid.*)

Coon<sub>i</sub> tin Coon<sub>i</sub> sễ thang.  
John think John will win  
'John thinks that John will win.'

Even more cross-linguistic variation with respect to the strength of the C-like tendency can be noticed when one considers Thai compared to Vietnamese and sees that a translation of (2.11a) is tolerated in one language but not the other.

(2.16) Thai (*Ibid.*)

John<sub>i</sub> choop John<sub>i</sub>.  
'John likes John.'

(2.17) Vietnamese (*Ibid.*)

\*John<sub>i</sub> thuong John<sub>i</sub>.  
'John likes John.'

Thirdly, where R-expressions can appear bound, they seem to discriminate between other R-expressions and pronouns with respect to what constitutes an acceptable binder. This can be illustrated by comparing the grammatical example in (2.14), above, with the ungrammatical (2.18), below.

(2.18) Thai (*Ibid.*)

\*Khaw<sub>i</sub> khit waa Coon<sub>i</sub> chalaat.  
He think that John smart  
'He thinks that John smart.'

This has prompted Lasnik to argue that Principle C should be divided into two separate statements. One, C<sub>1</sub>, which may or may not hold depending on a language-specific parametric setting and a second, C<sub>2</sub>, that is claimed to be a universal.

(2.19) *Principles C<sub>1</sub> and C<sub>2</sub>* (*Ibid.*)

*Principle C<sub>1</sub>*: An R-expression must be R-free.

*Principle C<sub>2</sub>*: An R-expression must be pronoun free.



However, despite the fact that Lasnik’s ‘split-C’ strategy gets empirical coverage for examples (2.14) – (2.18), the analysis appears inadequate for at least three reasons.

Firstly, it provides no explanation for why  $C_1$  can at times be violated in English, Classical Greek, and probably every language under certain contextual conditions, as in (2.12) and (2.13). Once the latter fact is admitted, there is once again no way of distinguishing a language like Thai from a language like English.

Secondly, Lasnik’s purported solution to the problem posed by the data in (2.14)–(2.18) suggests no explanation for why R-expressions and Anaphors do not actually appear in free variation even in, say, Thai. For while (2.14) and (2.16) may be more acceptable to Thai speakers than (2.11a) and (2.11c) are to English speakers, I assume that it could be clearly shown that (2.14) and (2.16) are highly dispreferred constructions in Thai as well, all things being equal.

In these ways,  $C_1$  is too strong for languages like English and to suppose that there is no  $C_1$ -type force whatsoever at work in languages like Thai seems incorrect as well.

A third issue has been raised. Namely, there is evidence that  $C_2$  is also not a universal, contrary to Lasnik’s claim.

(2.20) (Evans, 1980)

- a. Everyone has finally realized that Oscar is incompetent.
- b. Even he<sub>*i*</sub> has finally realized that Oscar<sub>*i*</sub> is incompetent.

(2.21) Chinese (Huang, 2000, 29)

Ta<sub>*i*</sub> you zai gan zhe xiaozi<sub>*i*</sub> yiguan gan de shi.  
3SG again DUR do this guy always do REL thing  
‘He’s doing just what the guy always does.’

For these reasons, the idea that pragmatic factors can at the very least interfere with whatever condition, principle, or force that is effecting patterns such as those illustrated in (2.11) is an idea that has been widely, if not universally, accepted. Chomsky formulates the sentiment as a general discourse principle.

(2.22) *Chomsky's general discourse principle* (Chomsky, 1981, 1982)

- a. Avoid repetition of R-expressions, except when conditions warrant.
- b. When conditions warrant, repeat R-expressions.

The general discourse principle is, in itself, just an admission that the challenges to Principle C noted above need to be explained away by making reference some sort of ‘competition’ between discursive, pragmatic factors and configurational or morphosyntactic ones. Others working in the generative grammar framework have gone further and concluded that Principle C ought not to be considered a part of the binding theory at all,<sup>6</sup> thus suggesting that the C-like effects we see across languages are likely due to pragmatic, discursive tendencies that are probably grounded in general principles of human and non-human behavior and are in no way specific to the language faculty.

Those within the generative grammar tradition have been generally very unwilling, however, to make similar conclusions about the remaining two principles of BT, Principles A and B.

### 2.2.2 Pronouns

Turning first to Principle B, the following example illustrates its supposed effects.

(2.23) John<sub>i</sub> likes him<sub>\*i/j</sub>.

Modern English poses few counterexamples to Principle B, though certain dialects tolerate violations more than others. Most tolerate (2.24) and many will tolerate (2.25) and often (2.26), for example.

(2.24) John<sub>i</sub> wrapped a blanket around him<sub>i</sub>.

(2.25) John<sub>i</sub> needs to get him<sub>i</sub> a shave.

(2.26) John<sub>i</sub> is going to buy him<sub>i</sub> a razor.

---

<sup>6</sup>Cf., e.g., Reinhart (1983, 1986), Grodzinsky & Reinhart (1993) , or Reinhart & Reuland (1993).

Moreover, it seems clear that there is semantic discrimination with respect to what kinds of predicates can manifest Principle B violations in idiolects that allow them. For example, while English dialects that allow (2.25) and (2.26) are quite common worldwide, no dialect to my knowledge will tolerate (2.27) or (2.28), below.

(2.27) \*John<sub>i</sub> needs to give him<sub>i</sub> a shave.

(2.28) \*John<sub>i</sub> is going to send him<sub>i</sub> some flowers.

What is more, cross-linguistically, Principle B meets many more blatant counterexamples.

First of all, as pointed out in Levinson (2000) et al., there are a considerable number of languages which appear to simply do without specialized words or morphemes that encode reflexivity. Examples include not only earlier dialects of English (in particular Old English) (Visser, 1963; Mitchell, 1985; Keenan, 2000) but also Australian languages like Guugu Yimithirr (Dixon, 1980), Austronesian languages such as Fijian (Dixon, 1988) and Tahitian (Tryon, 1970), as well as quite a few pidgins and creoles, e.g., Palenquero, Guadeloupe, KiNubi, and others (Carden & Stewart, 1988, 1987). In such languages, pronouns are used in bound A-positions and can (and, in certain contexts, must) be interpreted reflexively.

(2.29) Old English (Faltz, 1985, 19)

Swa hwa swa<sub>i</sub> eadmedaþ hine<sub>i/j</sub>...  
 Whoever humiliates him  
 ‘Whoever humiliates him(self)...’

(2.30) Tahitian (Tryon, 1970, 97)

’ua ha’opohe ’oia<sub>i</sub> ’iana<sub>i/j</sub>.  
 was kill he him  
 ‘He killed him(self).’

Secondly, some languages permit Principle B violations even though their lexical inventories contain some element that could arguably be considered a bona fide reflexive. In such languages, there is systematic overlap between

pronouns and reflexives in locally bound environments, which standard BT does not predict for. Examples might include languages such as (certain dialects of) Middle English (Visser, 1963; Faltz, 1985, et al.), Haitian Creole (Carden & Stewart, 1988) and, according to Levinson (2000, 339), perhaps as many half of all other creole languages.

(2.31) Haitian Creole (northern dialect) (Levinson, 2000, 338)

Emile<sub>i</sub> dwe ede (tèt-a-)li<sub>i</sub>.  
 Emile should help (head-of-)him  
 ‘Emile should help himself.’

Finally, it is well known that many Germanic, Romance, and other languages do not use (or at least do not have to use) reflexives to refer to local person antecedents, where by “local person antecedents” I mean antecedents with which person and number agreement implies coreference, i.e., first- or second-person antecedents.

(2.32) German

Du liebst dich (selbst).  
 ‘You love yourself.’

All these examples show that Principle B-like patterns can certainly not be described as universals without a great deal of qualification. This suggests that, much as was shown for the Principle C-like pattern, the B-like pattern might be most accurately described as a cross-linguistic tendency rather than a manifestation of a principle of Universal Grammar.

### 2.2.3 Anaphors

Finally, Principle A predicts the strictly local distribution of reflexives.

- (2.33) a. John<sub>i</sub> likes himself<sub>i/\*j</sub>.  
 b. John<sub>i</sub> thinks that Bill<sub>j</sub> likes himself<sub>\*i/j/\*k</sub>.

As with Principles C and B, Principle A seems to be fairly well regarded in English, though there too counterexamples can be found.

(2.34) (Huang, 2000, 95)

That everyone but herself<sub>i</sub> can play the viola depressed Mary<sub>i</sub>.

However – just as with Principles B and C – more obvious problems emerge when a wider range of cross-linguistic data is considered.

In the first place, the binding of reflexives can be sensitive to factors beyond just configurational relations. The fact that an A-B-C-type analysis will do nothing to explain the contrast between (2.35a) and (2.35b), below, has been used to illustrate this point.

(2.35) (Postal, 1971, 193)

- a. John talked to Mary about herself.
- b. \*John talked about Mary to herself.

Moreover, there are languages in which non-configurational factors seem to play a much more important role in governing the distribution of reflexives than configurational ones do. Examples include languages with nominative anaphors, such as Hungarian (Kiss, 1985), Malagasy (Randriamasimanana, 1996), and Modern Greek.

(2.36) Modern Greek (Everaert & Anagnostopoulou, 1997)

O eaftos tu<sub>i</sub> tu aresi tu Petru<sub>i</sub>.  
The self NOM his CL-DAT like-3SG the Petro-DAT  
'Himself pleases Petro.'

There seems no way that any version of BT could account for the grammaticality of (2.36), which contains nominative anaphor in (unbound) subject position and an antecedent in (locally bound) object position.

Such examples have prompted the search for semantically-based explanations for the distribution of reflexives; Fillmore (1968), Jackendoff (1972, 1990), Giorgi (1984, 1991), Wilkins (1988), and Grimshaw (1990) have all advocated the idea that the distribution of reflexives is (partially) regulated by a thematic condition, which makes reference to a thematic hierarchy.

(2.37) *Thematic Hierarchy Condition* (THC) (Jackendoff, 1972, 148)

A reflexive may not be higher on the Thematic Hierarchy than its antecedent.

(2.38) *Thematic Hierarchy* (Grimshaw, 1990, 24)

Agent > Experiencer > Goal/Source/Location > Theme

In (2.35a), the reflexive bears the  $\theta$ -role ‘Theme’ while the antecedent, *Mary*, is a ‘Goal’. On the other hand, in (2.35b) *herself* is the Goal while *Mary* is now the Theme. Therefore, since Goal outranks Theme on the Thematic Hierarchy, (2.35a) will satisfy the THC whereas (2.35b) will violate it, thus explaining why (2.35a) is acceptable and (2.35b) is out. Similarly, the thematic approach correctly predicts that (2.36) is acceptable, since in the antecedent is thematically more prominent than the anaphor (since the anaphor has the  $\theta$ -role of Theme and the antecedent is an Experiencer). Moreover, the thematic-based analysis would also correctly predict that, in contrast to (2.36), (2.39), below, is unacceptable.

(2.39) Modern Greek (Everaert & Anagnostopoulou, 1997)

\*O eaftos tu<sub>i</sub> ton xtipise ton Petru<sub>i</sub>.

The self-NOM his CL-ACC hit-3SG the Petro-ACC

‘Himself hit Petro.’

Example (2.39) is a simple transitive verb where the subject is an Agent and the verb a Theme, therefore in this case the THC is not satisfied.

Of course, it can be easily noticed that a purely thematic analysis would run into obvious problems as well.

(2.40) \*Himself pleases John.

But the examples above at least show that Principle A of BT does not provide an entirely adequate explanation for the distribution of so-called Anaphors and the evidence suggests that whatever force is responsible for Principle A-like effects like those in (2.33) is a force that ‘competes’, in a sense, with other forces, both across and within languages.

A second challenge facing the standard BT analysis of ‘Anaphors’ pertains to expressions like Icelandic *sig* and Chinese *ziji*. Such expressions are among the large number of examples often referred to as ‘long-distance Anaphors’ (or ‘long distance reflexives’) for precisely the reason that they may act as reflexives in locally bound environments but also systematically violate Principle A of BT.

(2.41) Icelandic (Sigurðsson, 1990)

Jón<sub>i</sub> segir að María<sub>j</sub> elski sig<sub>i/j</sub>.  
Jon says that Maria loves-SBJV self  
'Jon says that Maria loves him/herself.'

(2.42) Chinese (Huang, 2000, 94)

Xiaoaming<sub>i</sub> yiwei Xiaohua<sub>j</sub> zhidao Xiaolin<sub>k</sub> xihuan ziji<sub>i/k/j</sub>.  
Xiaoaming think Xiaohua knows Xiaolin like self  
'Xiaoming thinks Xiaohua knows that Xiaolin likes him/herself.'

Of any single challenge to any of the three Binding Principles, the problem of LDAs is perhaps the one which has earned the most space in the literature and there have been quite a few different proposals about how to address the problem generally and how to get a handle on the great amount of cross-linguistic variation which has been shown to exist with respect to what kind of expressions can be LDAs, how 'long distance' they can be, and what other kinds of restrictions they are subject to.

I will not even attempt to summarize the various approaches here other than to say that almost all of them either rely on the notion of movement at Logical Form (LF) or involve some recharacterization of the notion of 'Anaphor' and/or the notion of Accessible subject/SUBJECT and/or the notion of Governing Category.<sup>7</sup>

---

<sup>7</sup>This remark is made with six different types of approaches in mind, though there are others:

The *bound pronominal hypothesis* is the hypothesis that LDAs are not Anaphors, but rather pronouns and has been advocated in one form or another by Bouchard (1984), Sells (1987), and Pollard & Sag (1992).

The *pronominal anaphor hypothesis* treats LDAs as both pronoun *and* Anaphor, cf. e.g., Chomsky (1982, 78), Mohanan (1982), or Thráinsson (1991).

The *expansion hypothesis* expands the definition of a GC and has drawn support from Huang (1983), Wang & Stilings (1984), and Battistella & Xu (1990).

The *parameterization hypothesis* heralded by Manzini & Wexler (1987) parameterizes the definition of GC and is discussed briefly below.

The *relativization hypothesis* of Progovac (1992, 1993) involves restating the conditions for a binding and domain involves a redefinition of an accessible subject/SUBJECT so as to allow this definition to be different for LDAs than it is for strictly local reflexives.

Finally the *movement hypothesis* states that LDAs *do* satisfy Principle A, but that the surface representation of a sentence can obscure this fact because Anaphors undergo movement at LF. Cf. Lebeaux (1983), Pica (1985), Chomsky (1986), Battistella (1989),

Two properties of LDAs which have caused challenges to almost all BT-related research on the matter are the following.

Firstly, LDAs often exhibit strong *tendencies* in their behavior both cross-linguistically and within individual languages. Two of the most widely cited examples – and they are often referred to wrongly in the literature as linguistic universals – are the tendency of LDAs to take only subjects as antecedents and the tendency of LDAs to be morphologically simplex.<sup>8</sup>

(2.43) Marathi (Wali, 1989, 83)

Minine<sub>i</sub> Vinulaa<sub>j</sub> kaḷavle ki aapaṅ<sub>i/\*j</sub> turungaāt aahot.  
 Mini-ERG Vinu-DAT informed that self prison-LOC was  
 ‘Mini informed Vinu that he was in prison.’

(2.44) Chinese (Huang, 2000, 96)

Xiaoming<sub>i</sub> yiwei Xioahua<sub>j</sub> xihuan ziji<sub>i/j</sub>/taziji<sub>\*i/j</sub>.  
 ‘Xiaoming thinks Xioahua likes him(self).’

Some BT-related accounts are silent about the subject orientation and/or morphological simplicity of LDAs and thus do not predict for results like the ones shown in (2.43) and/or (2.44). Others are designed in such a way

---

Cole et al. (1990), Huang & Tang (1991), Katada (1991), Reinhart & Reuland (1991), et al.

The strategies mentioned above might be summarized as follows.

	recharacterizes Anaphor	recharacterizes accessible SUBJECT	recharacterizes governing category
Bound-pronominal hypothesis	+	–	–
Relativization hypothesis	–	+	–
Parameterization hypothesis	–	–	+
Pronominal-anaphor hypothesis	+	+	–
Expansion hypothesis	–	+	+
Movement hypothesis	–	–	–

It is beyond my scope to further discuss these approaches here, but for a beautiful summary and criticism of each of these approaches, cf. Huang (1994, 79-112) or Huang (2000, 101-126), who argues that all are inadequate by virtue of the fact that they can capture universals, but never universal *tendencies*, and this is a point I will take up later.

<sup>8</sup>Cf., e.g., Faltz (1985), Pica (1985, 1987), Reinhart & Reuland (1993), or Burzio (1998) for examples of the counterfactual claim that “when anaphors are complex expressions, they are universally local, whereas the long-distance type is universally simplex ... [and] subject-oriented (can be bound only by a subject).” (Reinhart & Reuland, 1993, 658-9)



that subject-orientation and morphological simplicity of LDAs follows from the mechanics of the analysis (in particular Progovac’s relativization-based account (Progovac, 1992, 1993) and most movement-oriented analyses.)

However, Huang (2000, 93-130), *inter alia*, has shown quite convincingly that the two properties in question are not universals. Cf. the morphologically complex *zibun-zisin* in (2.45) or the non-subject-oriented *sibi* in (2.46), below.

(2.45) Japanese, (Hara, 2002, 74)

John<sub>i</sub>-ga Mary<sub>j</sub>-ni Mike<sub>k</sub>-ga zibun-zisin<sub>i/\*j/k</sub>-o  
 John-NOM Mary-DAT Mike-NOM zibun-zisin-ACC  
 seme-ta-koto-o tuge-ta.  
 blame-PST-COMP-ACC tell-PST  
 ‘John told Mary that Mike blamed him(self).’

(2.46) Latin (Benedicto, 1991)

A Caesare<sub>i</sub> ualde liberaliter inuitor sibi<sub>i</sub> ut  
 By Caesar-ABL very generously am invited self-DAT COMP  
 sim legatus.  
 be-SBJV legate-NOM  
 ‘I am invited most generously by Caesar to be on his staff.’

Thus, once again, we see that an adequate account of the relevant phenomena seems to require a middle-ground position between postulating universal restrictions and no restrictions at all, and this is a position which a standard principle-based account cannot fit itself into. Parameterizing the issues would seem to do little good, since – as has been shown, again by Huang (2000, 93-130), *inter alia* – LDAs do *tend* to be morphologically simplex and subject-oriented, even in languages where there is no strict requirement that they be so.

A second difficulty which LDAs pose to virtually any GB-based theory of binding is the non-complementarity which they sometimes exhibit with pronouns.

LDAs are often limited with respect to how ‘long-distance’ they can actually go. Manzini & Wexler (1987), have attempted to explain this type discrimination by parameterizing the notion of Governing Category.

(2.47) *Governing Category Parameter* (Manzini & Wexler, 1987)

$\gamma$  is a GC for  $\alpha$  iff  $\gamma$  is the minimal category that contains  $\alpha$ , a governor for  $\alpha$ , and:

- a. has a subject, or
- b. has an inflection, or
- c. has a tense, or
- d. has an referential tense, or
- e. has a root tense.

M&W note an apparent implicational universal across languages involving these domain-types and claim that the universal can be nicely captured by assuming that the parameter choices obey a subset condition such that, for Anaphors, larger domains imply smaller domains and, for pronouns, the smaller domains imply the larger ones.

(2.48) *Subset hypothesis for parameter values*

- a. for Anaphors:  $La \subset Lb \subset Lc \subset Ld \subset Le$
- b. for pronouns:  $Le \subset Ld \subset Lc \subset Lb \subset La$

Per the subset hypothesis, if a particular LDA<sup>9</sup> in a particular language will tolerate being bound in, say, an embedded indicative clause, then it will always tolerate binding out of an embedded subjunctive, infinitival, or small clause. The implication works in the opposite direction for pronouns. Such an account, however, still predicts for the strict complementary distribution of pronouns and Anaphors, since whatever the GC is for a particular NP, pronouns must be free therein.

However – just as the prediction of the complementary distribution of pronouns and Anaphors predicted by Principles A and B is not borne out with examples like Middle English and Haitian Creole, where pronouns can show up where reflexives can too – languages with LDAs often exhibit the same type of non-complementarity in non-locally bound environments, so that there may be systematic overlap between pronouns and Anaphors either in certain types of embedded clauses (e.g., Icelandic subjunctives), or in virtually any type of embedded clause (e.g., Chinese).

---

<sup>9</sup>Cf. M&W's *Lexical Parameterization Hypothesis*: each anaphoric expression has its own parameter value, i.e., its own binding domain.

(2.49) Icelandic (Sigurðsson, 1990)

Jón<sub>i</sub> segir að María<sub>j</sub> elski sig<sub>i</sub>/hann<sub>i</sub>.  
Jon says that Maria loves-SBJV self  
'Jon says that Maria loves him/herself.'

This type of non-complementary distribution outside the (traditionally defined) local domain troubles the M&W-style treatment of LDAs just as non-complementarity within the local domain troubles the standard BT approach to local anaphora.

Burzio (1998) has suggested how M&W's parameterization idea can be nicely recast in the Optimality Theory framework of Prince & Smolensky (1993) and how – because Optimality Theory employs violable constraints which may be of comparable strength and thus allow for optionality – the possibility that a grammar could allow for the kind of non-complementarity shown in (2.49) can be left open. In particular, he proposes an Optimal Antecedent Hierarchy, whereby constraints are formulated to militate against LDAs and ranked universally so as to predict, for example, that an Anaphor that may be bound out of an indicative will always be bindable out of a subjunctive embedded clause as well, though the opposite is not necessarily the case.

(2.50) *Optimal Antecedent Hierarchy*

a. Subject of:

Indicative  $\gg$  Subjunctive  $\gg$  Infinitive  $\gg$  Small clause  $\gg$  NP

b. \*NP<sub>i</sub> ... [<sub>α</sub> NP ... SE<sub>i</sub> ...] ( $\alpha$  = clause)

The content of the Optimal Antecedent Hierarchy in turn participates in a Prominence Hierarchy, where the degree of prominence is “determined jointly by thematic role, discourse factors, and the semantic content of the inflection in the manner of [the Optimal Antecedent Hierarchy]” (Burzio, 1998, 100).

(2.51) *Optimal Prominence*

NP<sub>i</sub><sup>*p*</sup> ... SE<sub>i</sub>  $\gg$  NP<sub>i</sub><sup>*p*-1</sup> ... SE<sub>i</sub>  $\gg$  ... (*p* = prominence)

In this way, as Burzio puts it, “LD anaphora are thus possible so long as a loss in locality of the interpretive relation is offset by a gain in the prominence of the antecedent.” (*Ibid.*)

Burzio’s account, like M&W’s, can nicely capture cases where SE anaphora exhibit differential distribution based on configurational guidelines. One important class of cases are those LDAs which are not only *allowed* to appear long-distance in a certain type of embedded clause, but are *mandatory* in those clause-types, since pronouns are ungrammatical.

(2.52) Icelandic (Thráinsson, 1991, 51, 53)

Petur<sub>i</sub> bað Jens<sub>j</sub> að raka sig<sub>i/j</sub>/\*hann<sub>i/j</sub>.  
 Jon asked Jens that shave-INF SE/him  
 ‘Petur asked Jens to shave him/himself.’

On the other hand, because constraints related to thematic role, discourse factors and other factors that compete to determine prominence are never spelled out explicitly, Burzio’s account, as it stands, is unequipped to describe patterns like those in Chinese, Korean, and Japanese (Huang, 1994, et al.) and, according to Sigurðsson (1990), Old Icelandic, where, it appears, there is little evidence for configurational sensitivity of the kind represented by the Optimal Antecedent Hierarchy, since, in these languages, LDAs can appear in almost any type of embedded clause. Without any specific mention of *how* the notion of locality competes with other factors to determine prominence, the Prominence Hierarchy is little more than the Optimal Antecedent Hierarchy, plus some ‘general discourse principle’ (like the one Chomsky suggested to address Principle C counterexamples) that allows unspecified discourse factors to override syntactic stipulations.

Relatedly, the approach, as stated, misses the generalization pointed out by O’Connor (1993), Stirling (1993), and Levinson (2000) that there are semantic differences between LDAs and pronouns in contexts where both can appear, usually involving a contrast between a logophoric and non-logophoric reading.

In Chapter 6, I will suggest an approach that I believe can serve to address some of these issues. Presently, though, I turn to a major alternative to the standard BT within the generative tradition due to Reinhart & Reuland (1991, et al.).

## 2.3 Semantic Approaches to Binding Phenomena

The reflexivity framework of Reinhart & Reuland (1991, 1993, 1995) represents a radical alternative to the standard BT analysis discussed in the previous section. In their framework, there is no longer a simple distinction between Anaphors and pronominals. Rather, anaphoric NPs are classified into three groups according to two semantic properties: reflexivity and referential dependence. The new distinction is specifically aimed at differentiating what Bouchard (1984) was once led to call “true reflexives” such as English *himself* or Icelandic *sjalfan sig* and “false reflexives” like Icelandic *sig* or Italian, French, and Spanish *se*. R&R call the former type SELF anaphors and the latter SE anaphors, where the distribution of the relevant semantic properties among SELF anaphors, SE anaphors, and pronouns is as below.

(2.53)

	SELF	SE	Pro
<i>Reflexivizing function</i>	+	–	–
<i>Referentially independent</i>	–	–	+

The property of referential independence is essentially the one assumed in the Chomskyan BT framework for R-expressions, but R&R take pronouns to have the property as well. The  $\pm$  value for a reflexivizing function indicates whether or not the expression can reflexivize a predicate, where “reflexivize a predicate” means to indicate syntactically that two arguments of a predicate are conjoint (i.e., that the predicate is a reflexive predicate). Note that SELF anaphora are taken to be the only type of anaphor that has this property. Importantly, however, not all reflexive predicates need to be reflexivized. Rather, some predicates – specifically those which cannot take any object distinct in reference from the subject – are taken to be *intrinsically reflexive*.

The drawing of distinctions between SE and SELF anaphors and between intrinsically reflexive and intrinsically non-reflexive predicates is partly based on the need to capture a difference in licensing conditions for SE- and SELF-type expressions, for there are at times observable differences in the distributional behavior of the two, which, it seems, inevitably depend on the lexical properties of the relevant verb.

(2.54) Dutch (Reinhart & Reuland, 1993)

- a. Max schammt zich.
- b. \*Max schammt zichzelf.  
'Max is ashamed.'

(2.55) a. \*Max bewondert zich.  
b. Max bewondert zichzelf.  
'Max admires himself.'

The patterns above trouble traditional accounts of BT for the obvious reasons; none of the sentences in (2.54) or (2.55) violates Principle A and thus should be acceptable. On R&R's account, the discrepancy is taken to arise from the fact that *schammen* is an intrinsically reflexive verb that need not be reflexivized and – in the interest of avoiding redundancy – never will be. As such, (2.54b) will always be bad. On the other hand, *bewonderen* is not an intrinsically reflexive verb and to get a reflexive interpretation for a case like (2.55), it will need to be reflexivized; (2.55a) is out then, because SE anaphors cannot reflexivize predicates. This line of reasoning is formally stated in R&R's version of Principles A and B.

(2.56) *Binding Principles A and B*

*Principle A:* A reflexive marked syntactic predicate is reflexive.

*Principle B:* A reflexive semantic predicate is reflexive marked.

(2.57) *Definitions*

- a. The *syntactic predicate* formed of a head *P* is *P*, all of *P*'s syntactic arguments, and an external (subject) argument of *P*.
- b. The *syntactic arguments* of *P* are the projections assigned a  $\theta$ -role or Case by *P*.
- c. The *semantic predicate* formed of *P* is *P* and all its arguments at the relevant semantic level.
- d. A predicate formed of *P* is *reflexive* iff two of its arguments are coindexed.
- e. A predicate formed of *P* is *reflexive marked* iff either *P* is lexically reflexive or one of *P*'s arguments is a SELF anaphor.

Principle B will rule out (2.55a), which is a reflexive semantic predicate that is not reflexive marked. Principle B does *not* rule out the coindexation in (2.58), below.

(2.58)  $\text{Max}_i$  schammt  $\text{hem}_i$ .

The predicate in (2.58) is intrinsically reflexive and thus satisfies Principle B. Both arguments are coindexed and therefore Principle A is also satisfied. In order to rule out such coindexation, R&R propose a condition on ‘A-chains’ which serves to further regulate the distribution of pronouns.

(2.59) *A-chain* (Reinhart & Reuland, 1993, 693)

An *A-chain* is any sequence of coindexation that is headed by an A-position and satisfies antecedent government.

(2.60) *General condition on A-chains* (Reinhart & Reuland, 1993, 696)

A maximal A-chain  $(a_1, \dots, a_n)$  contains exactly one link –  $a_1$  – that is both referentially independent and Case marked.

By virtue of the general condition on A-chains, (2.58) is now ruled out, since a chain formed between *Max* and *hem* would contain two referentially independent, Case marked elements. Furthermore, the condition on A-chains induces a requirement that the referentially independent element of the chain must c-command the referentially dependent one.

(2.61) \*Himself<sub>i</sub> loves  $\text{Max}_i$ .

Example (2.61) violates the general condition on A-chains since it is headed by a link that is not referentially independent. The unique link that is referentially independent (and Case marked) in this case is *Max*. But *Max* is at the tail of the chain, not the head.

R&R’s approach represents an important paradigm shift in the treatment of binding phenomena in that it is the first formal treatment of binding phenomena to recognize and exploit the distinction invoked by, e.g., Farmer & Harnish (1987) and Levinson (1991, et al.), regarding whether or not an anaphor and its antecedent coarguments of a single predicate. By allowing that distinction to serve as the nucleus of the theory rather than relying on notions like ‘government’ and ‘GC’, R&R’s account makes empirical improvements on standard BT as well. Consider:

- (2.62) a. John<sub>i</sub> wrapped a blanket around him<sub>i</sub>.  
b. John<sub>i</sub> wrapped a blanket around himself<sub>i</sub>.

We have seen how the Principle B of standard BT will incorrectly rule out (2.62a) (since the GC here is the whole sentence and *him* is not free). However, (2.62) does not contain a semantically reflexive predicate and thus R&R's Principle B is satisfied in both (2.62a) and (2.62b). Moreover, the SELF anaphor in (2.62b) does not reflexive-mark the relevant predicate since *himself* is not a syntactic argument of the head, *wrapped*, and thus Principle A is not violated either.

Despite such advantages, however, the approach R&R advocate is met by a number a serious challenges.

Most importantly, the main empirical prediction of analysis – that only intrinsically non-reflexive predicates should get reflexive marked and that those predicates must always be marked if they are to be interpreted reflexively – is just not supported by the majority of cross-linguistic data; most languages simply do not exhibit the same type of pattern observed in Dutch.

Firstly, there are languages like English which mark reflexivity with *-self* morphemes even when the verb in question is intrinsically reflexive.

- (2.63) John behaved himself.

One could dismiss this example on the basis of fact that, lacking a SE anaphor, English has no choice but to use a SELF anaphor, for if it used a pronoun it would violate Condition B and thus the non-redundancy issue is not relevant here. But even then, languages which *do* use pronouns as the objects of reflexive predicates – either optionally or because they lack reflexives altogether – will still pose counterexamples, cf. (2.29)-(2.31), above, or (2.64), below.

- (2.64) Frisian (Everaert, 1991, 94)

Hy skammet him.

‘He is ashamed.’

What is more, even in the majority of languages that do utilize SE anaphora, reflexive marking via *self*-type morphemes is non-obligatory, even with predicates that are not intrinsically reflexive.



(2.65) German

Johann liebt sich (selbst).

‘Johann loves himself.’

(2.66) Korean (Cole et al., 1990, 18)

Chelswu-nun (caki)-casin-ul sarangha-n-ta.

Chelswu-TOP (SELF)-SE-ACC love-PRES-DECL

‘Chelswu loves himself.’

In addition to this, one finds – especially in languages with SE anaphora or with verbal reflexives – cases of ‘double-marking’ in the literature, where a verbal reflexive can take a reflexive as a complement, as in (2.67), or where SE-type clitic anaphora can co-occur with a pronoun (as in (2.68)) or a SELF anaphor, like in (2.69).

(2.67) Japanese (Aikawa, 1993, 76)

John-ga zibun-(zisin)-o ziko hihansita.

John-NOM zibun(-zisin)-ACC self criticized

‘John criticized himself.’

(2.68) Padovano (Lidz, 1996, 43)

Gianni se varda lu.

Gianni SE saw him.

‘Gianni saw himself.’

(2.69) Spanish (Huang, 2000, 164)

Ana se vio (a sí misma).

Ana SE saw SELF

‘Ana saw herself.’

It would be difficult to explain cases of double-marking within R&R’s program, given the economy/nonredundancy guidelines that militate against marking things that are already marked.

Finally, just as with the standard BT, R&R’s theory of reflexivity does not provide an account of LDAs<sup>10</sup> and makes no reference to thematic relations which, we saw, were especially relevant for languages like Greek or Malagasy, which seem to follow binding patterns which must be explained in thematic terms.

---

<sup>10</sup>R&R handle LDAs with an account based on movement at LF, cf. (Reinhart & Reuland, 1991, 291, 301-308).

## 2.4 Summary

Above I have sketched two competing theories of binding in the generative grammar tradition, one configurationally-based analysis of Chomsky and second a semantically-based approach of Reinhart & Reuland. We saw that while Chomsky's Principle C has been largely marginalized or abandoned in these frameworks, Principles A and B – either Chomsky's versions or Reinhart & Reuland's – have been largely maintained to be universal and innate principles of grammar. However, much of the evidence cited above suggests that, much like the supposed effects of Principle C, Principles A and B might also be more accurately described as cross-linguistic tendencies as opposed to cross-linguistic universals, and thus might best be explained in terms of something other than Universal Grammar. In the following chapter I turn to a much different view on binding phenomena, namely the work of Levinson (1991, 2000, et al.), who advocates a pragmatic approach to binding phenomena partially motivated by counterexamples to more traditional forms of binding theory like the ones discussed above.

## Chapter 3

# Pragmatic Approaches to Binding Phenomena

### 3.1 Introduction

It is perhaps evident from the previous chapter that wide acceptance has been given to various approaches to binding phenomena which rest solely on configurational notions such as government and binding and/or semantic notions like reflexivity. But the position that syntactic and/or semantic factors are the only ones relevant to an adequate analysis of binding phenomena is not universally held and, in the last fifteen years or so, references to pragmatic considerations such as ambiguity avoidance and conversational implicature in analyses of intrasentential anaphora have grown more common and less subtle.

Chomsky himself has never advocated any systematic pragmatic theory of anaphora nor does he ever suggest, to my knowledge, that universal binding patterns relate in any way to pragmatic considerations, though one does encounter the occasional reference to pragmatic principles such the ‘general discourse principle’, cited in Chapter 1, which can permit for the relaxation of Principle C under certain (unnamed and unspecified) conditions.

To be sure, pragmatic-sounding notions like ‘economy’ and ‘least effort’ have been invoked in Chomsky’s Minimalist program (Chomsky, 1995), wherein derivational and representational economy are considered to be the functional source of grammatical principles such as Shortest Move, Procrastinate, and so on. However, Chomsky has made no claim to the effect that

these notions of economy are more general instantiations of pragmatic tendencies. On the contrary, Chomsky considers the aforementioned principles to be peculiar to the human language faculty, so such a relationship is actually explicitly denied.

However, strong proposals have been made, most notably by Levinson (1987b, 1991, 2000, et al.) as well as Huang (1994, 2000, et al.) to the effect that the consideration of pragmatic factors is essential to an explanatorily adequate analysis of the set of universal patterns and tendencies now commonly known as binding phenomena.

There are several reasons for believing that, insofar as these phenomena are concerned, pragmatic factors deserve a great deal more credit than they have traditionally been given.

Firstly, it seems almost beyond doubt that the behavior of intrasentential anaphora and the behavior of intersentential anaphora are in some way related.

- (3.1) a. John entered the pub.  
b. ?Then John ordered a drink.

But syntactic or semantic theories of binding are typically silent about questions surrounding unbound anaphora and seem to treat the sentence boundary as a sacred border that serves to demarcate separate domains of syntax or semantics on the one hand and pragmatics on the other. There seems to be little justification for this attitude, especially when we consider languages like Old English, Guugu Yimithirr, or any other language which lacks reflexive expressions or uses them optionally in locally bound environments. In such languages, any distinction between the strategies used for interpreting bound anaphora and those used for resolving unbound anaphora would seem to be a very fuzzy one.

What is perhaps an even more convincing body of evidence in favor of a pragmatic approach to anaphoric binding behavior is the data gotten from examining diachronic change in various languages and the general trends such change tends to follow. Binding patterns in individual languages have been shown to change considerably over time and the gradual diachronic evolution of a language like English – which, over the last 1200 years or so, has gone from a language that patently lacked reflexives to one which gradually developed reflexive pronouns, which eventually became mandatory in most dialects – causes obvious challenges to any principle-based account

of binding phenomena as well as to many parameter-based learning models. A pragmatics-based account is, at the very least, more naturally suited to deal with the large amount of optionality within certain languages that one finds, especially in early stages of diachronic development. Moreover, if one accepts the idea that pragmatic effects could ‘grammaticalize’ over time, then pragmatic approaches to binding can potentially offer predictions about the general direction of diachronic development that semantic or syntactic approaches seem to be in no position to make.

Few if any proposals have ever been put forth that suggest that pragmatics is solely responsible for binding phenomena. Rather, any attempt at a pragmatic analysis of binding phenomena has typically included reference to autonomous syntactic, semantic and pragmatic levels of explanation. The claim, then, is simply that an adequate analysis of intrasentential anaphora cannot ignore pragmatic factors, and the main question of interest is exactly what roles these factors play and what part of the labor they share in governing the phenomena in question.

The most radical claims to date begin with Levinson (1987b, 1991, 2000), who suggests that pragmatics is central to the issue of binding phenomena and advocates an analysis largely aimed explaining Binding Principle-like effects in terms of either pragmatic inferences or ‘fossils’ thereof, where a ‘fossil’ is some rule or constraint or semantic feature resulting from a pragmatic inference having grammaticalized.

## 3.2 Gricean Pragmatics

The most significant precursor to Levinson’s ideas was the work in pragmatics and philosophy done by Grice (1957, 1975, 1978, 1989). Grice develops a theory of ‘non-natural’ or non-literal meaning, the idea being 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*.

(3.2) *Grice's Cooperative Principle* (Grice, 1989, 26)

“Make your contribution as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.”

(3.3) *Grice's Conversational Maxims* (*Ibid.*, 26, 27)

*Quality*

- a. Do not say what you believe to be false.
- b. Do not say that for which you lack adequate evidence.

*Quantity*

- a. Make your contribution as informative as is required.
- b. Do not make your contribution more informative than is required.

*Relation*

Be relevant.

*Manner*

- a. Avoid obscurity.
- b. Avoid ambiguity.
- c. Be brief (avoid unnecessary prolixity).
- d. 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):

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

While no compositional analysis of an utterance like (3.4) 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. In Grice's terms, (3.4) generates a *particularized* conversational implicature, as it is basically context dependent since the phrase *a man* cannot normally be counted on to induce the implicature '*not Mr. Jones*'. In contrast, *generalized* conversational implicatures are those implicatures which an expression generally conveys in the absence of information to the contrary, just as an utterance of the expression *some* generally conveys the unspoken caveat *but not all*, as in (3.5), below.

(3.5) Some of my friends smoke.

Levinson has suggested a way to apply Gricean ideas to the explanation of binding phenomena. His approach involves one of a handful of reductionist accounts of Grice's conversational maxims heralded in the literature of the 'radical pragmatics' tradition.

### 3.3 Radical Pragmatics

The central hypothesis behind the 'radical pragmatics' tradition is that various linguistic phenomena which have previously been treated as being governed by syntactic or semantic factors actually fall under the scope of pragmatics.<sup>1</sup> The work done in the area of radical pragmatics has been done mostly by various 'neo-Griceans', especially Horn (1984, et al.) and Levinson (1987b,a, et al.), who have made advances and revisions in Grice's original model for (the calculation of) conversational implicatures, normally in the form of reductionist accounts and/or enrichments and specifications of the schemas under which the various parts of the Cooperative Principle interact.

The inspiration for much of the work in neo-Gricean pragmatics can be traced back to Zipf (1949), who argues that the development of a language is largely a result of two opposing forces – the 'force of unification' and the 'force of diversification'. The force of unification is a sort of 'principle of least effort' seen from the speaker's perspective which serves to minimize vocabulary, whereas the force of diversification has the opposite effect and is related with the hearer's drive to minimize his own efforts.

Much in the spirit of Zipf's idea(s), Horn (1984) proposes a bipartite model in which the three Gricean maxims of Quantity, Relation, Manner are reduced to two basic principles, (Q)uantity and (R)elation.

---

<sup>1</sup>This is basically a paraphrase of Cole (1981, *Introd.*).

(3.6) *Horn's Q- and R-principles*

*Q-principle*: Make your contribution sufficient; say as much as you can (given R).

*R-principle*: Make your contribution necessary; say no more than you must (given Q).

Horn argues that the Gricean mechanism for pragmatic implicature can be derived from the interaction of the Q- and R-principles in accordance with a 'division of pragmatic labor', which, in effect, allows obedience to the R-principle unless a contrastive linguistic expression is used that induces a Q-implicature to the effect that the R-principle is inapplicable.<sup>2</sup>

(3.7) *Horn's division of pragmatic labor*

The use of a marked (relatively complex and/or prolix) expression when a corresponding unmarked (simpler, less 'effortful') alternative expression is available tends to be interpreted as conveying a marked message (one which the unmarked alternatives would not or could not have conveyed).

Levinson (1987b,a) airs grievances with Horn's bipartite reductionist model, the basic objection being that that model lacks any distinction between semantic or informational economy on the one hand (semantically general expressions ostensibly being, for a speaker, more economical than semantically specific ones) and structural economy on the other (less complex expressions being preferred to more complex ones, by the speaker). With the desire for drawing such a distinction in mind, Levinson revises Horn's program in a way that distinguishes between those pragmatic principles that pertain to the surface complexity of an expression and those which relate to the informational content of the expression to yield a system with three, not two, pragmatic principles – the I-, Q-, and M-principles, which as a whole constitute Levinson's theory of generalized conversational implicatures (GCIs). Crucially, each principle involves not only a speaker-oriented maxim, but also a hearer-oriented corollary.

The I-principle is, as with Horn's R-principle, a maxim of (constrained) minimization for a speaker, while for a hearer it is a maxim of (constrained) maximization.

---

<sup>2</sup>Though cf. Horn (1989, 192-203) for a discussion of an algorithm for determining whether Q or R takes precedence in a particular discourse context.



(3.8) *I-principle* (Levinson, 2000, 114, 115)

*Speaker maxim:*

‘Say as little as necessary’, i.e., produce the minimal linguistic information sufficient to achieve your communicational ends (bearing the Q-principle in mind.)

*Hearer corollary:*

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, unless the speaker has broken the maxim of minimization by using a marked or prolix expression. Specifically:

- a. Assume the richest temporal, causal, and referential connections between described situations or events, consistent with what is taken for granted.
- b. Assume that stereotypical relations obtain between referents and events, unless this is inconsistent with (a).
- c. Avoid interpretations that multiply entities referred to (assume referential parsimony); specifically, prefer coreferential readings of reduced NPs (e.g., pronouns or zeros).
- d. Assume the existence or actuality of what a sentence is ‘about’ if that is consistent with what is taken for granted.

I-implicatures are the result of the ‘amplification’ mentioned in the I-hearer corollary. 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. On the other hand, the speaker maxim directs one to use semantically general statements wherever semantically less general statements are unnecessary. The main pragmatic effect of the I-principle is to induce the hearer to select an interpretation (out of a number of possible interpretations) that best comports with the most ‘stereotypical’ state-of-affairs, given his knowledge of the world.

The general interpretational strategy represented by the I-principle has been given credit for a wide range of linguistic/discursive phenomena; examples include conjunction buttressing (e.g., Atlas & Levinson (1981)), bridging inferences (e.g., Clark & Haviland (1977)) and indirect speech acts (e.g.,

Searle (1975)).<sup>3</sup>

(3.9) John pushed Bill. He fell.

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

(3.10) 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 the speaker maxim of the I-principle and, for a hearer, a maxim of minimization essentially serving to curb the amplification licensed by the I-principle's hearer corollary.

(3.11) *Q-principle* (Levinson, 2000, 76)

*Speaker maxim:* Do not provide a statement that is informationally weaker than your knowledge of the world allows, unless providing an informationally stronger statement would contravene the I-principle. Specifically, select the informationally strongest paradigmatic alternative that is consistent with the facts.

*Hearer corollary:* Take it that the speaker made the strongest statement consistent with what he knows and therefore that:

a. If the speaker asserted  $A(W)$ , where  $A$  is a sentence frame and  $W$  an informationally weaker expression than  $S$ , and the contrastive expressions  $\langle S, W \rangle$  form a Horn-scale (in the prototype case, such that  $A(S)$  entails  $A(W)$ ),<sup>4</sup> then one can infer that  $Know(\neg S)$ , i.e., that the

---

<sup>3</sup>But see Levinson (2000, 117) for a richer list of examples and references.

<sup>4</sup>Technically, a Horn-scale, per Horn (1972), is defined as follows

(3.12) *Horn-scale*

$\langle S, W \rangle$  forms a Horn-scale only if

a.  $A(S)$  entails  $A(W)$  for some arbitrary sentence frame  $A$

b.  $S$  and  $W$  are equally lexicalized

c.  $S$  and  $W$  are 'about' the same semantic relation or from the same semantic field.

speaker knows the stronger statement,  $S$ , is false.

b. If the speaker asserted  $A(W)$  and  $A(W)$  fails to entail an embedded sentence  $P$ , which a stronger statement  $A(S)$  *would* entail, and  $\{S, W\}$  form a contrast-set,<sup>5</sup> then infer  $\neg Know(P)$ , i.e., 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  $S$  and  $W$  form a Horn scale or contrast set.

The Q-principle will do the work of Grice's Quantity(a) and will induce conversational implicatures that arise from a Horn-scales and/or contrast sets. Illustrative examples include *scalar* implicatures coerced, per hearer corollary (a), by quantifiers in the appropriate type of opposition with one another.

(3.13) Some of my friends smoke.

Q-implicature: *Not all* of my friends smoke.

A second group of Q-phenomena, *clausal* implicatures, effected by hearer corollary (b), involve contrasts between expressions which entail a certain embedded sentence and expressions which do not, cf., e.g.,  $\langle think, know \rangle$ .

(3.14) John thinks Mary loves him.

Q-implicature: Mary *might not* love John.

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

---

The importance of restricting Q-implicatures to 'Horn-scale pairs' can be appreciated by observing that, from a sentence like *John thinks Mary loves him*, 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 (3.12a)).

<sup>5</sup>Cf. Levinson (2000, 76-111) for discussion of various types of 'Q-contrasts' which appear to reliably induce implicatures where no Horn-scale exists. E.g., under the GCI picture, *John tried to convince Mary that he was fool* could Q-implicate '...did not convince...' by virtue of a contrast-set  $\{try, succeed\}$ , despite the fact that succeeding does not entail having tried.

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, unlike the Q- and I-principles which refer to semantic informativeness, the M-principle relates to surface complexity.

(3.15) *M-principle* (Levinson, 2000, 136, 137)

*Speaker maxim*: Indicate an abnormal, nonstereotypical situation by using a marked expression that contrasts with one you would normally use to describe the corresponding normal, stereotypical situation.

*Hearer corollary*: What is said in an abnormal way indicates an abnormal situation, or marked messages indicate marked situations, specifically: Where a speaker has uttered a marked expression  $M$  to say  $p$  and there is some unmarked expression  $U$  which the speaker could have used in the same sentence frame instead and  $U$  and  $M$  have the same semantic denotation, then where  $U$  would have I-implicated the stereotypical or more specific subset  $d \subseteq D$ , the marked expression  $M$  implicates the complement of the denotation of  $d$ , i.e.,  $\bar{d} \subseteq D$ .

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 I-implicature that would typically be associated with the minimal sufficient expression.

The M-principle is meant to represent Horn's division of pragmatic labor whereby "unmarked forms tend to be used for unmarked situations and marked forms for marked situations" (Horn, 1984, 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*/*\*furiousity*) – 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, cf., e.g., Atlas & Levinson (1981) or Horn (1984). One classic example from McCawley (1978):

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

Here, a simple lexical causative like the one in (3.16a), can describe a run-of-the-mill act of homicide, whereas the productive causative in (3.16b) – though unacceptable for describing stereotypical murder, manslaughter, etc. – 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 (3.16b) – and generally any M-implicature – is one which coerces an interpretation of non-stereotypicality due the use of a marked expression despite the availability of an unmarked one.

### 3.4 Radical Pragmatics and Anaphora

Levinson proposes that some of the Binding Principles of Chomsky's BT follow from patterns of preferred interpretation effected by GCIs.

Because 'preferred interpretations' are, in principle, defeasible, they do not typically render some interpretation impossible for some form, in contrast with supposedly inviolable syntactic conditions like the Binding Principles.<sup>6</sup> Thus, if we find a pattern of anaphoric interpretation in some language that does not appear to be at all defeasible – cf. e.g., \**John<sub>i</sub> is pleased with him<sub>i</sub>* – we are justified in believing that the interpretations which constitute that pattern are not merely 'preferred' and are thus not patterns which can be fully explained in terms of GCIs.

However, Levinson argues that we are still fully entitled to suspect that any 'indefeasible preference' or tenet of grammar in a language that does not noticeably *conflict* with the well-known defeasible Gricean patterns might be a manifestation of those patterns. In particular, he hypothesizes that what are, in some languages, seemingly indefeasible, syntactic regulations (like the Binding Principles) 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 that language.

---

<sup>6</sup>In fact, Levinson claims that the property of defeasibility is "*the litmus test*" (Levinson, 2000, Ch. 4, fn. 6, my italics) for whether something is a product of grammatical stipulation or preferred interpretation.

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. Evidence of this type exists, in particular, languages like Old English, early Haitian Creole, and so on, mentioned in Chapter 2, which pose challenges to standard BT exactly because they lack morphological means of encoding reflexivity altogether and use pronouns reflexively, thus disobeying Condition B systematically and obeying Principle A only vacuously.

Levinson's so-called B-then-A account is a story of how, assuming the three principles of GCI theory to be at work, the effects of Principles 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 & 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' (cf. Levinson (2000, 286-327) for discussion), though they are perhaps better 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. As is discussed below, the approach has some advantages over a principle-based account of binding phenomena not only in that it is a reductionist account, but also because it impressively avoids or addresses some of the problems that BT and other theories have struggled with.

Levinson's 'B-then-A' or 'B-first' account takes as a starting point a pre-existing anaphoric pattern in which something like Chomsky's Principle 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 (*Ibid.*, 329-330). 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 & Harnish, 1987, 557) .

(3.17) *Disjoint Reference Presumption (DRP)*

Arguments of a predicate are intended to be disjoint, unless marked otherwise.

Levinson, following Carden & Stewart (1988) , identifies 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.

(3.18) *Stages 1-3 (Levinson, 2000, 339)*

*Stage 1*: no encoded reflexives; plain pronouns used reflexively

*Stage 2*: gradual emergence of morphological reflexives (based on, e.g., body-part expressions or emphatics) with a clausemate, subject antecedent condition, coexisting but encroaching on upon the use of an ordinary pronouns

*Stage 3*: loss of reflexive use of ordinary pronouns

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” (*Ibid.*, 374) can be used to M-implicate the reversal of the DRP, i.e., locally conjoint reference.

Levinson cites a considerable number of examples of languages which, as mentioned in Chapter 2, appear to do without specialized words or morphemes that encode reflexivity – including Old English, Australian languages like Guugu Yimithirr, Austronesian languages such as Fijian, as well as quite a few pidgins and creoles, e.g., 18th century Haitian Creole, Palenquero, Guadeloupe, KiNubi, and others (*Ibid.*, 338-341). In these cases, reflexivity

is typically expressed by a piece of, say, detransitivizing verbal morphology (like Guugu Yimithirr), or stressed or emphasized pronoun (like Haitian Creole), or unreduced object pronoun (like Fijian), which “encourages a coreferential reading” (*Ibid.*, 336), but does not guarantee it. Levinson (2000, 350), following Faltz (1985), eventually makes the general claim that “nearly all reflexives ultimately arise from emphatic or stressed pronouns”.

A further example is English itself, though not its modern form. Specifically, as noted in Chapter 2, evidence from Old English (cf. Visser (1963, 420-439) and Mitchell (1985, 115-189)) 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 perhaps preferably interpreted as reflexive, did not necessarily induce a locally conjoint interpretation.

(3.19) Old English (Mitchell, 1985, 115)

Moyses<sub>i</sub>, se ðe wæs Gode<sub>j</sub> sua weorð ðæt he<sub>i</sub> oft wið  
 Moses, he who was to-God so dear that he often with  
 hine selfne<sub>j</sub> spræc.  
 him self spoke

‘Moses was so dear to God that he often spoke with God himself.’

According to Levinson’s account, in Stage 1 languages, the anti-locality (i.e., Principle B-type) effects for pronouns and the locality (i.e., Principle A-type) effects for ‘proto-reflexives’, e.g., 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.

(3.20) Old English<sup>7</sup> (Visser, 1963, 433)

- a. He<sub>i</sub> ofsticode hine<sub>j>i</sub>.
  - b. He<sub>i</sub> ofsticode hine selfne<sub>i>j</sub>.
- ‘He stabbed him(self).’

---

<sup>7</sup>The grammaticality judgements below reflect the discussion in Levinson (2000, 341), citing Visser (1963), who, in turn cites Sweet (1882).



Because the hearer corollary of the M-principle directs a hearer to interpret the marked expression *hine selfe* as a speaker’s deliberate avoidance of the “stereotypical associations and I-implicatures of” the available, unmarked expression *hie*, (3.20b) will be viewed as a deliberate avoidance of (3.20a), and thus (3.20b) will get whatever interpretation (3.20a) normally does *not* get. Per the DRP, (3.20a) gets a disjoint reading. Thus, (3.20b) gets a coreferential reading, per the hearer corollary of the M-principle. Of course, both of these preferences, are, as yet, defeasible, cf. the ambiguity of (3.20a) and (3.20b).

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 might include (according to Carden & Stewart (1988)) Martinique Creole, Mauritian Creole, Bislama, and presumably various dialects of Old and Middle English.

A language has reached Stage 3 when the aforementioned emphatic expressions can fairly be said to have grammaticalized into legitimate reflexive markers. Levinson claims that, for English, we might draw this line at the point where the emphatic expression lost its inflection, at the transition from Old to Middle English. (cf. *hine selfne* > *hie selfe*).

(3.21) Middle English<sup>8</sup> (Visser, 1963, 421)

- a. He<sub>i</sub> forseoð hie<sub>j</sub>><sub>i</sub>.
  - b. He<sub>i</sub> forseoð hie selfe<sub>i/??j</sub>.
- ‘He despises him(self).’

According to Levinson’s GCI based analysis, the (preferred) disjoint interpretation of (3.21a) is due now 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  $\langle hie\ selfe, hie \rangle - hie\ selfe$  being, according

---

<sup>8</sup>The judgements below again reflect Levinson (2000, 341 and fn. 69/Ch. 4) citing Visser (1963, 439).

to Levinson, the more informative expression.<sup>9</sup>

As for Principle C-type effects, in languages at each stage, Levinson derives the effect 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:

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

According to the hearer corollary of the M-principle, 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 (3.22b) does not mean what he could have expressed with (3.22a), since he is avoiding (3.22a) at a cost to himself, presumably to avoid exactly those I-implicatures that (3.22a) would typically effect (especially coreference).

Moreover, in Stage 3 languages, the Condition C-like effect is reinforced by 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  $\langle \textit{himself}, \textit{John} \rangle$ , *himself* still being the more informative expression, according to Levinson.

In summary, the (albeit defeasible) Condition A- and B-like effects are, for Levinson, viewed as symptoms of the DRP, M-implicatures and Q-implicatures, while the Condition C-type effects are attributed to M-, and sometimes also Q-, implicatures.

### LDA's and Logophoricity

Levinson has further argued that his GCI-based approach to anaphoric paradigms can be also extended to offer empirical coverage for LDA's.

The linchpin of his analysis is the claim – following, O'Connor (1993) and Stirling (1993), et al. – that an 'Anaphor' "carries additional information which the pronoun is not marked for" (Levinson, 2000, 314) and that

---

<sup>9</sup>Based on notions of informativeness due to Bar-Hillel & Carnap (1952), Levinson (2000, 273, 274) advocates the idea that a reflexive is more informative than a non-reflexive pronoun because a coreferential interpretation introduces a smaller number of entities into the domain of discourse compared to a non-coreferential one and thus is the less general, more specific interpretation.

LDAs are “always referentially dependent and always logophoric.”, where “logophoric” means that the expression carries a “marked deictic perspective ... having something to do with emphatic contrast, empathy, or protagonist’s perspective, subjective point of view, and so on.” (*Ibid.*, 312, 347)

Thus, while what Levinson (and Chomsky) call an Anaphor – like Icelandic *sig* – may potentially have the same reference as pronoun, it always contrasts semantically with a pronoun on some other level since the Anaphor will carry perspectival information and the pronoun will not.

Levinson proposes that the process of becoming ‘marked for logophoricity’ is – just like the process of becoming marked for reflexivity – a diachronic phenomenon that can be divided into the three stages. In total then, Levinson’s Stages 1-3 can be summarized as follows.

(3.23) *Stages 1-3 (revised)* (Levinson, 2000, 347, 348)

*Stage 1: No reflexives; disjoint interpretations of core arguments are preferred*

a. Sentences of the form ‘*John hit him*’ will I-implicate disjoint reference, via the DRP; ad hoc means of M-implicating conjoint reference can be found in intonational stress, emphatics, and so on.

b. Furthermore, in certain constructions, particularly in embedded constructions like ‘*John thinks that Mary loves him*’, the same ad hoc means can be used to M-implicate some marked deictic perspective, rather than coercing an M-implicature toward local conjoint reference.

*Stage 2: Emphatic core-arguments may be preferably conjoint*

a. The exceptional coreferential interpretation of a sentence like ‘*John hit him*’ is now regularly reinforced by the use of a marked or emphatic pronoun: ‘*John hit him-emph*’ M-implicates a locally conjoint interpretation.

b. An established usage of this sort reinforces the contrast in such a way that an unmarked pronoun as in ‘*John hit him*’ I-implicates disjoint reference more strongly.

c. Other contrasts besides reference may also be M-implicated by marked or emphatic pronouns. One possibility being point of view contrast or ‘logophoricity’, especially relevant outside the scope of the DRP. So, ‘*John thinks Mary loves him*’ expresses the unmarked deictic point of view, whereas ‘*John thinks Mary loves him-emph*’ M-implicates

a marked point of view.

*Stage 3: Emphatics become reflexives*

An established system emerges: grammaticalized reflexives that encode necessary referential dependence within some domain have evolved and thus:

a. *In coargument positions.* The pronoun in ‘*John hit him*’ is not only presumed disjoint by the DRP, uncanceled by any M-implicature, but also Q-implicates disjoint reference by the scale  $\langle \textit{Anaphor}, \textit{pronoun} \rangle$ . Hence the strong inference to disjoint core arguments from the use of a pronoun.

b. *Where antecedents and anaphoric expressions are not clausemate coarguments:* If the DRP does not apply or is overridden, the contrast in reference between a pronoun and ‘Anaphor’ is not the only possible one. Instead, the marked anaphor may indicate a marked point of view:

	<i>Anaphor</i>	<i>Pronoun</i>
<i>coreference</i>	+	±
<i>logophoricity</i>	+	±

Thus, the use of a pronoun will Q-implicate that the speaker is not in a position to use the Anaphor either because he intends disjoint reference or he intends non-logophoricity.

### 3.5 Summary

Levinson’s Stage 1-3 analysis provides an interesting alternative to traditional generative grammar approaches to binding not only because it obviates the postulation of innate principles of grammar by reducing those principles to artefacts of pragmatics, but also because it is in a good position to handle cases which have troubled traditional semantic or syntactic approaches, especially cases of languages which lack reflexives or exhibit widespread optionality with respect to the distribution of pronouns and reflexives and/or SE anaphora. Moreover, the approach is able to make some predictions about the direction of diachronic change – many of which seem to be in a broad sense correct when we look at the kind of diachronic development of a language like Old English into Modern Standard English, where an emphatic

suffix seems to have evolved from occasional conversational cue that marked contrast in reference or other dimension into a genuine reflexive marker that was mandatory in core-conjoint environments.

However, the locality of reflexives and the antilocality of pronouns in languages like Modern English are arguably no longer defeasible, thus the GCI-based analysis Levinson advocates is not quite strong enough to provide empirical coverage for such cases. Levinson claims in several places that such indefeasibility can be attributed to some process of grammaticalization whereby the GCIs of Old and Middle English have ‘frozen’ or ‘fossilized’ somehow. However, while the observations that a suffix like English *-self* took on a reflexive meaning in bound A-positions and that *self*-marked forms gradually replaced pronouns in those positions are both well supported, the actual mechanism of grammaticalization is left very much open in Levinson’s discussion.

Relatedly, the GCI-based picture would have difficulty explaining why reflexive marking would spread to cases that did not involve any type of potential ambiguity. For example, while a form like *himself* may have emerged as a contrastive alternative to *him*, there is no such explanation for why a form like *myself* came to be. The same could be said for cases of discriminatory *self*-marking, like that of Dutch, discussed in Chapter 2, whereby objects of intrinsically reflexive predicates must be unmarked, while other objects may not be. No semantic differentiation actually results from these types of markings, and thus any M-implicatures that are predicted would presumably point in the wrong direction.

In the remaining chapters, I wish to suggest that recent advances in the field of Optimality Theory and OT-based models of language learning can be used to recast Levinson’s account in a way that both preserves some of the fortunate results and adds a great deal of clarity to the story about grammaticalization that the account relies so heavily on.

# Chapter 4

## Optimality, Superoptimality & Anaphora

### 4.1 Introduction

In the present chapter I present a pragmatic approach to the explanation of basic binding patterns that is meant to improve on the work of Levinson (2000, et al.), discussed in Chapter 3.

The approach is set in the Optimality Theory (OT) framework of Prince & Smolensky (1993) and relies heavily on Blutner's (2000b) ideas of bidirectional optimization. In particular, I will suggest an explanation of how some of the basic anaphoric paradigms – especially those exemplified by what Levinson called 'Stage 1' and 'Stage 2' languages, discussed in Chapter 3 – could show up as the result of the interaction of a few commonsense constraints plus the effects of bidirectional optimization.

In Chapter 5 I will introduce the ideas of Zeevat and Jäger regarding the role of interpretational bias in language comprehension and diachronic change (Zeevat & Jäger, 2002; Zeevat, 2002) as well as the work of (Cable, 2002) and (Jäger, 2003a), both of whom use advances in OT research due to Boersma (1998) to give a more precise account of grammaticalization.

Finally, in Chapter 6, I will try to show that these accounts can be further improved upon and extended to yield an evolutionary account of marking strategies that predicts for the marked-form-for-marked-meaning pattern noted by Horn (1984) and Blutner (2000b), *inter alia*, and can provide a precise formal account of the evolution of a Stage 1 language into a Stage 3


language and can thus, when appropriate, mimic the empirical coverage of traditional binding theories as well.

In the last decade or so, research in generative linguistics has witnessed a sharp increase in the frequency of references to notions of economy and other concepts that necessarily involve reference to some sort of conflict or competition.<sup>1</sup> The notion of economy and the notion of competition go hand in hand, for judging an expression or operation to be economical can only be done if the expression or operation being evaluated can be compared to various alternatives.

The OT of Prince & Smolensky (1993) is a framework in which the competition between linguistic entities is a central notion. In OT, a certain *input* gets associated with a multitude of possible *outputs* or *candidates* (this set is known as GEN). Each candidate is then evaluated with respect to a series of ranked, violable constraints (collectively known as EVAL). 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, where the definition of relative harmony is as below.

(4.1) *Relative Harmony* (Prince & Smolensky 1993)

Relative to a constraint hierarchy,  $H$ , a candidate,  $\alpha$ , is more harmonic than a candidate,  $\beta$ , (write:  $\alpha \succ_H \beta$ ), if  $\alpha$  ‘better-satisfies’  $H$ , where “better satisfies  $H$ ” means that  $\alpha$  commits less violations of a constraint  $C$  than  $\beta$  does, where  $C$  is the highest ranked constraint in  $H$  with respect to which  $\alpha$  and  $\beta$  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 results of an evaluation procedure are typically represented via *tableaux*, which depict the constraint-violation tallies yielded from cross-referencing the constraints with the candidates (for some given input). The candidates appear on the left-hand vertical axis while the constraints are above, horizontally. A ‘\*’ represents a violation, ‘\*!’ represents a fatal violation, and ‘’ represents an optimal candidate.

---

<sup>1</sup>Chomsky’s Minimalism (1995 et al.) and its reference to ‘Shortest Moves’ and so on is the most well known case. Reinhart & Reuland’s (1993) reference to an informal nonredundancy condition, discussed in Chapter 2, is just one more example.

(4.2)

<i>Input</i>	$C_1$	$C_2$	$C_3$
$\alpha$		*	**!
$\beta$		*	*
$\gamma$	*!		*

The OT framework was originally proposed as a theory of generative phonology and has subsequently been exploited in the fields of morphology and syntax (cf. Grimshaw (1997), Bresnan (1998), Burzio (1998), et al.). Recent work including but not limited to that of van der Does & de Hoop (1998), de Hoop & de Swart (2000), and Hendriks & de Hoop (2001) has applied OT to semantics. The major distinction between the first approaches to OT semantics and previous applications of OT to phonology, morphology, and syntax is that the semantically geared versions are interpretational, not generative, enterprises and thus the pertinent constraints judge candidate meanings with respect to input forms, not candidate forms with respect to input meanings.

## 4.2 Bidirectional OT

In other recent proposals, Blutner (2000b), Zeevat (2001), Jäger (2003a), et al. have all argued that *bidirectional optimization* – i.e., the combination of generative and interpretational optimization – is of central importance if we wish to apply OT to the semantics and pragmatics of natural language. With a generative dimension added to ‘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 interdependence of the two dimensions, for it is exactly this interdependence that is the major focus of the Grice and Levinson literature, as it is generally seen as the root cause of most conversational implicatures and as the main reason for the emergence of what Horn (1984) called the ‘division of pragmatic labor’, discussed in Chapter 3.

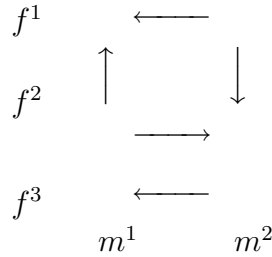
Where we write ‘ $\langle f, m \rangle$ ’ to stand for some form/meaning pair, we can write ‘ $\langle f', m \rangle \succ \langle f, m \rangle$ ’ to mean that, relative to  $m$ ,  $f'$  is more harmonic than  $f$  and ‘ $\langle f, m' \rangle \succ \langle f, m \rangle$ ’ to mean that, relative to  $f$ ,  $m'$  is more harmonic than  $m$ . The definition of bidirectional optimality is then straightforward.





These results (and the results of any Bi-OT analysis) can be represented in ‘arrow diagrams’, due to Dekker & van Rooy (2001), who note parallels between the Bi-OT literature and work in Game Theory.

(4.5)



Here, the horizontal arrows represent the interpretational preferences relative to the various forms, the arrows pointing to the left showing  $m^1$  to be most harmonic for  $f^1$  and  $f^3$ , and the arrow pointing to the right signifying that the optimal candidate for  $f^2$  is  $m^2$ . Likewise, the vertical arrows show the generative preferences relative to the relevant meanings. Here,  $f^1$  is the optimal candidate, given  $m^1$ , and  $f^2$  is optimal for  $m^2$ . The absence of any arrow selecting  $f^3$  means that  $f^3$  is blocked (i.e., blocked by another form, in this case,  $f^1$ ).

This formulation of bidirectional optimality enables one to model cases of *total blocking*, whereby some forms (e.g., *\*yesterday night*, *\*furiousity*) do not exist because other forms 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  $\langle f, m \rangle$  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, being that the primary, initial motivation for developing a bidirectional version of OT was the interest in capturing the Gricean and neo-Gricean results heralded in the radical pragmatics literature and tradition of Atlas & Levinson (1981), Horn (1984), et al., 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 *superoptimality* 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.

(4.6) *Bidirectional Optimality (weak version)*

A form/meaning pair,  $\langle f, m \rangle$  is bidirectionally optimal iff:

- q.* there is no distinct pair  $\langle f', m \rangle$  such that  $\langle f', m \rangle \succ \langle f, m \rangle$  and  $\langle f', m \rangle$  satisfies *i*.
- i.* there is no distinct pair  $\langle f, m' \rangle$  such that  $\langle f, m' \rangle \succ \langle f, m \rangle$  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 superoptimal, 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 superoptimal status only if there is a superior pair  $\langle f', m \rangle$  or  $\langle f, m' \rangle$  *and* the superior pair is itself superoptimal. 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 is included in the definition for satisfaction of the *q*-condition. However, as Jäger, who has explored the formal properties of superoptimal evaluation (Jäger, 2000), points out, the definition is not circular so long as we assume that the ' $\succ$ ' relation is a well-founded one.

Consider McCawley's example once again.

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

In order to say why the marked form in (4.7b) gets associated with a marked meaning, we need to say explicitly what marked forms and meanings are. In OT, constraints alone determine what is marked and what is not. We can suppose, then, that two constraints like the following might be at work.

(4.8) *Cause*: Interpret causatives directly.

*Econ*: Avoid productive, compound, or analytic expressions.

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

(4.9)

<i>direct</i>	<i>Econ</i>
☞kill	
cause to die	*!

<i>kill</i>	<i>Cause</i>
☞direct	
indirect	*!

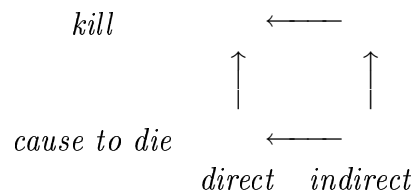
<i>indirect</i>	<i>Econ</i>
☞kill	
cause to die	*!

<i>cause to die</i>	<i>Cause</i>
☞direct	
indirect	*!

And hence:

(4.10)



While the diagrams above involve only one strongly bidirectionally optimal pair, both the pair  $\langle \textit{kill}, \textit{direct} \rangle$  and the pair  $\langle \textit{cause to die}, \textit{indirect} \rangle$  are superoptimal. Specifically, though the pair  $\langle \textit{cause to die}, \textit{indirect} \rangle$  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 superoptimal, exactly because there is no superoptimal pair  $\langle \textit{cause to die}, m \rangle$  such that  $\langle \textit{cause to die}, m \rangle \succ \langle \textit{cause to die}, \textit{indirect} \rangle$  and there is no superoptimal pair  $\langle f, \textit{indirect} \rangle$  such that  $\langle f, \textit{indirect} \rangle$  is more harmonic than  $\langle \textit{cause to die}, \textit{indirect} \rangle$ .

In this way, Blutner's idea lets us use a small set of commonsense constraints do the work of Levinson's I- and Q-principles and harvest the effects of the M-principle without further stipulation, by virtue of the mechanics of Bi-OT.

### 4.3 A Pragmatic/OT Approach to Binding Phenomena

Applying this idea specifically to the question of binding phenomena will allow us to capture some of the empirical results of Levinson's GCI-based analysis of those phenomena while at the same time reducing the ontological commitments of the analysis by obviating the M-principle.<sup>2</sup> In Chapters 5 and 6, I will propose that the relationship between (Bi-)OT and existing theories of learning and grammaticalization gives the approach even further advantages.

To begin with, let us assume that on the hearer-side of an evaluation, the following constraints represent 'hearer economy'.

(4.11)  $f$ : Do not license  $f$ -features.

$\varphi$ : Do not license  $\varphi$ -features.

$\delta$ : Do not license  $\delta$ -features.

The constraints  $f$ ,  $\varphi$  and  $\delta$  all militate against the licensing of features where by 'licensing' I mean specifying features for an interpretation that were left underspecified in the relevant expression being interpreted.<sup>3</sup> Thus, these constraints essentially represent a hearer-preference for maximally specific outputs with respect to the  $f$ -,  $\varphi$ -, and  $\delta$ -features of NPs, where the distribution of  $f$ -,  $\varphi$ -, and  $\delta$ -features is as below:<sup>4</sup>

---

<sup>2</sup>The idea of applying the notion of weak bidirectional optimality to recast Levinson's account of binding phenomena is due to Blutner as well. Though never published and slightly different from the account I advocate below, the same general idea was proposed on at least three occasions, all of which are memorialized in electronically available notes: Blutner (2000a), Blutner (2001), and Blutner (2002).

<sup>3</sup>The constraints  $f$  and  $\varphi$  are not novel, but are proposed by Buchwald et al. (2002) in an analysis of discourse pronouns (though there they go by the names ' $S\Phi^k$ ' and ' $S\Phi_k$ ', respectively, for reasons on which I will not elaborate). It is not lost on me that better names for these constraints might be ' $*\varphi$ ', ' $*f$ ', etc., since the constraints direct a hearer to *avoid* licensing the feature, but this might also be confusing as the constraint does not militate against the presence of the feature, but rather toward it, so I keep the notation as is.

<sup>4</sup>I borrow the term ' $f$ -feature' from Buchwald et al. (2002) and it can be considered as basically the semantic content of the actual name or description – perhaps something like the Predicate attribute in Bresnan & Kaplan's LFG (1982). The term ' $\varphi$ -feature' refers, as usual, to agreement features. The term ' $\delta$ -features' is my own term and is just meant

(4.12)

	Full	Pro	SE	$\emptyset$
<i>f</i> -features	+	-	-	-
$\varphi$ -features	+	+	-	-
$\delta$ -features	+	+	+	-

I will assume throughout that generative and interpretational constraints directly interact<sup>5</sup> and thus the constraints *f*,  $\varphi$ , and  $\delta$  will have important effects for the generative evaluation procedure. In particular, they will effectively act as a speaker mandate which says that where a feature is present (in the input), it should be linguistically expressed (in the output) and in this way they will be doing some of the same work as the speaker maxim of Levinson's Q-principle.<sup>6</sup> Because possession of *f*-features implies the possession of  $\varphi$ -features, which in turn entails the possession of  $\delta$ -features, a universal ranking  $\delta \gg \varphi \gg f$  can be assumed.

To represent preferences with respect to core-disjoint versus core-conjoint interpretations of predicate arguments, I assume for now that the following constraints are consulted.

- (4.13) \**dis*: Coarguments of a predicate are conjoint.  
 \**co*: Coarguments of a predicate are disjoint.

These constraints are obviously in direct conflict with one another. I will stipulate for the time being that a universal ranking \**co*  $\gg$  \**dis* exists and this will yield a DRP-like effect. By stipulating such a ranking, of course, the constraint \**dis* becomes irrelevant and could be left out altogether, but I introduce it here to illustrate a point that I will return to later, namely that a ranking schema like \**co*  $\gg$  \**dis* – and thus the DRP-like effect associated with it – will eventually follow from the present analysis without stipulation.

I assume that a final constraint can represent ‘speaker economy’.

- (4.14) \**Struct*: Avoid morphological structure.

I will show below how the constraints above, in combination with bidirectional optimization will enable us to recast Levinson's account in an OT framework.

---

to distinguish unpronounced elements from so-called SE anaphors, and can perhaps be thought of as referring to case or perhaps just some detransitivizing property.

<sup>5</sup>This assumption originates with Wilson (2001) and has been called *Recoverability OT* by Buchwald et al. (2002).

<sup>6</sup>As well as some of the same work as Burzio's *Optimal Agreement Hierarchy* (Burzio, 1998) or Wilson's *FtrFaith* (Wilson, 2001).

## 4.4 Some Applications: Simulating some reflexivizing strategies of ‘Stage 1’ languages

### 4.4.1 Introduction

The purpose of the remarks below is to show how the binding patterns of Levinson’s ‘Stage 1’ languages can be represented in the Bi-OT framework discussed above.

### 4.4.2 Old English

Recall that – per Visser (1963, 420-439) and Mitchell (1985, 115-189), et al. – 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*. OE *hine selfne* was an emphatic pronoun which, while it could be interpreted reflexively in a bound A-position, could also be interpreted non-reflexively and could appear unbound or even as a subject.

(4.15) Old English (Keenan, 2001, 8 (from Beowulf 960))

Uþe ic swiþor þæt ðu hine selfne geseon moste...  
How I wish that you him self seen be able-PST  
‘How I wish you could have seen him!’

Likewise, bare pronouns like OE *hine* and even Middle English *hie* and *hym* did not share the anti-locality restrictions their modern descendant.

(4.16) Middle English (Faltz, 1985, 19)

He<sub>i</sub> cladde hym<sub>i/j</sub> as a poure laborer.  
‘He dressed him(self) as a poor laborer.’

In terms of the constraints above, we can represent the OE pattern as follows. Inventorywise, we could represent the lack of null objects and SE anaphora via a high ranking for both  $\delta$  and  $\varphi$  relative to *\*Struct*. For simple transitive clauses, such a ranking pattern suggested would harvest the following generative results.<sup>7</sup>

---

<sup>7</sup>I note violations of a particular constraint only where one or more candidates differ from each other with respect to their performance.

(4.17)

$R(j, x) / R(j, j)$	$\emptyset$	$\varrho$	<i>*Struct</i>	<i>*co</i>	<i>*dis</i>
$\text{pro}$					
pro+emph			*!		
(SE)		*!			
( $\emptyset$ )	*!				

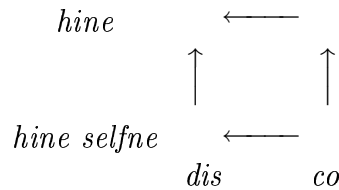
We see that the pronoun is the overwhelming favorite here, regardless of which meaning is intended. Likewise, due to the (stipulated) ranking of *\*co* and *\*dis*, we know that a core-disjoint interpretation will be the optimal one given either of the forms.

(4.18)

<i>Subj-Verb-pro(+emph)</i>	$\emptyset$	$\varrho$	<i>*Struct</i>	<i>*co</i>	<i>*dis</i>
R(j,x)					*
R(j,j)				*!	

We have:

(4.19)



Note that – just as with the  $\langle \textit{kill}, \textit{cause to die} \rangle$  case discussed above – we no longer need to stipulate the existence of a Levinsonian M-implicature to pair the dispreferred, emphatic pronoun with the dispreferred, conjoint interpretation, since, by definition,  $\langle \textit{hine selfne}, \textit{co} \rangle$  already satisfies the criteria for a superoptimal solution.<sup>8</sup>

<sup>8</sup>Note however, that this in no way commits us to the claim that  $\langle \textit{hine selfne}, \textit{co} \rangle$  is the only superoptimal pair that involves the expression *hine selfne*. As noted above, a locally conjoint interpretation was not the only one that was available, for there were potentially many other ways in which OE *selfne* could be used contrastively, i.e., many different dimensions besides the referential dimension, in which *hine selfne* could contrast with *hine*. This multitude of superoptimal possibilities could be produced only if we considered constraints like say *\*emphatic*, and so on, which would need to be of comparable strength with *\*co*, though for simplicity I leave consideration of these aside.



### 4.4.3 Pidgins and Creoles

Levinson (2000) – relying much on evidence provided by Carden & Stewart (1988, et al.) – notes that creole languages provide a very diverse reservoir of examples which seem to lack bona fide reflexives and thus, like Old English, would qualify as Stage 1 languages. Carden & Stewart and Levinson note creoles such as Arabic-based KiNubi, Spanish-based Palenqueno, French-based Guadeloupe, and others as examples of languages that have survived a long time without developing reflexive expressions. Moreover, claims Levinson (citing C&S as well as Corne (1988)), “perhaps half of all Creole lects ... allow the reflexive use of clausemate pronouns” (Levinson, 2000, 339); examples include Kriyol, Martinique Creole, Mauritian Creole, Bislama, Negerhollands, and modern Northern, as well as 18th century Haitian Creole. Haitian Creole, like a great number of other languages, expresses reflexivity with a compound of the form *body-part expression + pronoun*, though in certain modern dialects (and certainly in earlier dialects) the compound does not appear to have fully grammaticalized.

(4.20) Haitian Creole (northern dialect) (Levinson, 2000, 338)

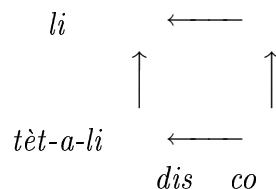
Emile<sub>i</sub> dwe ede li<sub>i</sub>.  
Emile should help him  
'Emile should help him/himself.'

(4.21) Emile<sub>i</sub> dwe ede tèt-a-li<sub>i</sub>.

Emile should help head-of-him  
'Emile should help himself.'

In the present framework, we can treat a language like early Haitian Creole – as well as any other language which displays similar patterns in which a body-part expression is combined with a pronoun to (eventually) form a reflexive marker – in exactly the same way we handled Old English. For if we take it that, as before, coarguments of a verb are preferably interpreted as disjoint (per the ranking schema *\*co* ≫ *\*dis*) and that *body-part expression + pronoun* compounds are dispreferred structurally compared to simple pronouns (by virtue of *\*Struct*) then we again see a case in which the simplex pronoun forms a strongly bioptimal pair with a locally disjoint interpretation but where a locally conjoint interpretation forms a weakly bioptimal pair with the dispreferred compound form.

(4.22)



As noted, large group of languages seem to fit into this category. Not all such languages can be fairly called Stage 1 languages, since, in some cases, the *body-part + pronoun* compound has effectively grammaticalized into a reflexive, but I will return to a discussion of grammaticalization in the following chapters.

#### 4.4.4 Australian and Austronesian languages

According to Levinson, Australian and Austronesian languages are two other groups of languages which often exhibit Stage 1 behavior. Many such languages again lack reflexives and employ some structurally marked alternative to a bare pronoun as a way of coercing a reflexive interpretation. Examples include the Austronesian languages Tahitian (Tryon, 1970) and Kilivilla (Senft, 1986; Levinson, 2000) and the Australian languages Guugu Yimithirr (Dixon, 1980), Nyawaygi (Dixon, 1983), Jiwari (Austin, 1987), and Gumbaynggir (Eades, 1983), the latter of which can express reflexivity with the nominal emphatic suffix *-w*, though the suffix is not mandatory for soliciting a reflexive reading.

(4.23) Gumbaynggir (Eades, 1983, 312)

Gua:du    bu:rwang    gula:na    magayu.  
3SG-ERG   paint-PST   3SG-ABS   red paint-INSTR  
'He painted him/himself with red paint.'

(4.24) Gua:du    bu:rwang    gula:naw                    magayu.

                  3SG-ERG   paint-PST   3SG-ABS-EMPH   red paint-INSTR  
'He painted himself with red paint.'

Such languages could be treated in exactly the same way that English and Haitian Creole were treated above; the marked form will form a weakly bioptimal pair with the dispreferred, core-conjoint interpretation.

More interesting, however, are a few cases in which the reflexivizing strategy is quite different. In Fijian, for example, according to Dixon (1988, 255-256), a verb with the transitive marker *-a* which lacks an overt object is interpreted as referring to a third-person singular object distinct from the subject. On the other hand, if coreference or reflexivity is intended, a full object pronoun (e.g., *'ea*, a third-person singular object) is required. And though a full pronoun does not necessarily require a conjoint interpretation, such interpretations are preferred.

(4.25) Fijian (Levinson, 1991, 135)

sa va'a-dodonu-ta'ini  $\emptyset$  o Mika.  
 ASP correct ART Mike  
 'Mike corrected him/her/it.'

(4.26) sa va'a-dodonu-ta'ini 'ea o Mika.  
 ASP correct 3SG-OBJ ART Mike  
 'Mike corrected himself/him.'

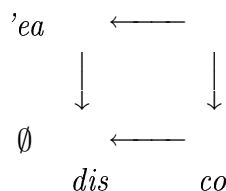
If we want to describe the local anaphoric pattern of Fijian discussed above in terms of the constraints proposed previously, we can represent, as usual, that locally disjoint reference is preferred to locally conjoint reference per the ranking  $*co \gg *dis$  and suppose that  $\delta$  is dominated by  $*Struct$ . We would have, say:

(4.27)

$R(j, x) / R(j, j)$	$*Struct$	$\delta$	$\varrho$	$*co$	$*dis$
pro	*!				
(pro+emph)	*!*				
(SE)	*!		*		
$\emptyset$		*			

In this case, it is  $\langle \emptyset, dis \rangle$  which forms the strong-bidirectionally optimal pair and  $\langle pro, co \rangle$  is the weak pair.

(4.28)



Thus it is the bare pronoun that is the marked form which can be expected to potentially solicit a core-conjoint reading.

### SE anaphora

Finally, one of the most common reflexivizing strategies is the employment of so-called SE anaphora

(4.29) Icelandic

Jón<sub>i</sub> elskar sig<sub>i</sub>.

Jón loves SE.

‘Jón loves himself.’

In terms of the present analysis, there is actually more than one way we might treat ‘Stage 1 SE anaphora’, depending on what one believes to be the origin of these expressions. (And, of course, the origins of SE anaphora in various languages might be quite different.)

Faltz (1985, 256-269) discusses the historical origins of SE anaphora and claims that one reasonable hypothesis is that the early ancestors of SE anaphora – perhaps, for example, (reconstructed) proto-Indo-European *\*s(w)-* – derived from stressed pronouns, similar to the way certain logophoric pronouns may have.<sup>9</sup>

If the hypothesis that PIE *\*s(w)-* was a stressed pronoun is correct, then our treatment of the PIE reflexivization strategy would be much like our treatment of OE and Haitian Creole. That is, the stressed pronoun would have formed a weakly optimal pair with the dispreferred, locally conjoint reading by virtue of some markedness constraint, (not *\*Struct*, but perhaps, say, Schwarzschild’s *AvoidF(ocus)* (Schwarzschild, 1999)) and we can expect that this form could have been used to indicate a contrast in reference with the normal pronoun, e.g., to indicate conjoint reference.

<sup>9</sup>Examples might include Igbo *yá*, Yoruba *oun*, and Lakhota *iye* (Faltz, 1985, 257-258).

A second hypothesis (discussed in Faltz (*Ibid.*, 259-262)) is that SE anaphora derived from first person pronouns<sup>10</sup> which generalized to uses for all persons, effectively dropping their person features (and apparently, in many cases, their number features as well). The opinion of many (e.g., Reinhart & Reuland (1993, et al.), discussed in Chapter 3) seems to be that SE anaphora are more appropriately analyzed as having *underspecified* their way to ‘referential dependence’ as opposed to specifying for it and being referentially dependent *by virtue of* their lack of  $\varphi$ -features, rather than the lack of  $\varphi$ -features being a “reflex” of their referential dependence, as Levinson suggests at one point (2000, 312).<sup>11</sup>

I have no personal views on the origin of SE anaphora, but if the idea that SE anaphora got their referential dependence through underspecification is in any way correct then we could alternatively attribute the markedness of such expressions to their offense with respect to the constraint  $\varphi$ , since a SE anaphor like PIE *\*s(w)-* would violate  $\varphi$  whereas a  $\varphi$ -feature endowed bare pronoun like PIE *\*tha* would not.

(4.30)

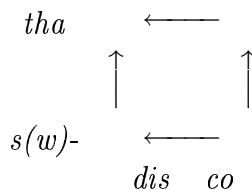
$R(j, x) / R(j, j)$	$\emptyset$	<i>*Struct</i>	$\varphi$	<i>*co</i>	<i>*dis</i>
$\text{pro}$		*			
pro+emph		**!			
SE		*	*!		
$\emptyset$	*!		*		

With this, and with our usual *\*co*  $\gg$  *\*dis* assumption in place, we have:

<sup>10</sup>Evidence for this path of evolutionary development might include the so called ‘reportive first person’ pronouns of Efik and Ewe (Faltz, 1985, 256-269), where in a sentence like *John says I love you*, the *I* could refer to John, not the speaker.

<sup>11</sup>Levinson (2000, fn 80, Ch. 4) tends to downplay the hypothesis that SE anaphora could have evolved from first person reportive pronouns. And it seems he has reason to; entertaining the hypothesis that SE anaphora took on reflexive meaning through underspecification would cause Levinson’s GCI-based account some real trouble. In particular, if SE-type expressions could no longer be considered specially marked or ‘heavy’ NPs then it would no longer be obvious why they could be said to be ‘more marked’ or ‘more informative’ than pronouns and hence the relevant M-implicatures, Horn-scales, and Q-implicatures would 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).

(4.31)



## 4.5 Summary

Above I've sketched a way to reconstruct Levinson's analysis of Stage 1 languages by letting Blutner's notion of weak optimality do the work of Levinson's M-principle.

Below, I wish to show how a Bi-OT-based account, since it is compatible with a powerful learning theory, can in turn lend itself to an elegant account of grammaticalization and can thus provide a precise description of mechanism for the transition between Stages 1-3 and beyond. Moreover, it can help deal with other issues that the account above, like Levinson's account, has trouble addressing, such as the spread of *self*-marking to local person arguments, which, just as they are not predicted by reference to M-implication, are not predicted by the notion of weak bidirectional optimality either (since *I hit me* will be interpreted optimally as meaning 'I hit myself', one would think).

## Chapter 5

# Bias, Stochastic Optimization, Gradual Learning & Grammaticalization

### 5.1 Introduction

The present chapter discusses the work of Zeevat & Jäger (2002) and Zeevat (2002), who suggest that the notion of interpretational bias based on statistical frequencies in language use can be used along with the idea of bidirectional optimization to explain how certain grammars come to possess certain marking rules. This idea has been further explored by Cable (2002) and Jäger (2003a), both of whom have combined the ideas of interpretational bias and bidirectionality with Boersma's stochastic OT and his Gradual Learning Algorithm (Boersma, 1998) to give formal accounts of the grammaticalization of marking strategies, in particular, marking strategies which follow the familiar marked-form-for-marked-meaning pattern noted by Horn (1984) and Blutner (2000b), *inter alia*. In my closing remarks to this chapter, I will suggest what I feel may be a slight improvement on the work of Cable and Jäger and then, in Chapter 6, I will argue that such an account can at least begin to provide an evolutionary explanation of the transition from Levinson's 'Stage 1' to 'Stage 3' and, more generally, can perhaps provide a functional, evolutionary explanation for (and hence an alternative to) Horn's division of pragmatic labor and Blutner's notion of weak bidirectional optimality.

In the previous chapter, I recast Levinson’s analysis of Stage 1 languages using a Bi-OT framework in place of the GCI-based approach that he advocates. I suggested that a Stage 1 language – basically any language which lacks reflexives and uses some device such as emphatic morphemes, body-part expressions, or unreduced or  $\varphi$ -feature impoverished pronouns to coerce locally conjoint interpretations – could be described as a language in which weak bidirectional optimality relations hold between the ‘marked’ expressions and dispreferred, locally conjoint interpretations in the same way that such a relation holds between an expression like *cause to die* and an interpretation of indirect causing of death. On such an account, both unmarked expressions (like a bare pronoun, in English) and marked ones (like a *self*-marked pronoun) would typically prefer locally disjoint interpretations, though the marked expressions could also be used for the purpose of soliciting some marked, especially conjoint, interpretation. We already know, however, that the interpretation of bare and *self*-marked pronouns is, at least in languages like Modern English, more than a matter of preferred interpretation and not likely a matter of a hearer’s conscious reflection on conversational principles. Conversational implicatures are characteristically overridable by explicit semantic information, but most would agree that this is just not the case for the interpretation of a sentences like *John loves himself* or *John liebt sich*.

We saw how Levinson distinguishes languages in which the reflexive interpretation of marked pronouns is generally a matter of implicature from those in which a bona fide reflexive has shown up and gradually become mandatory in locally conjoint environments by proposing that the former are Stage 1 or 2 languages and the latter are Stage 3 languages, where Stage 3 languages are those in which pragmatic preferences had grammaticalized somehow. But the mechanism for grammaticalization was, as noted, left largely open by Levinson.<sup>1</sup>

It is at least clear that a Stage 1-to-Stage 3 story (or anything like it) is *not* obviously compatible with one line of explanation, namely any parameter-based account whereby the resetting of a single parameter would effect a ‘binding shift’ whereby marked pronouns became strictly local and/or mandatory in local environments. Both Levinson (2000) and Keenan (2001)

---

<sup>1</sup>Huang (2000) also proposes a typological distinction between what he calls *syntactic languages* and *pragmatic languages*, though he ignores talk of diachrony altogether. He suggests too that the ‘syntacticness’ and ‘pragmaticness’ of any given language is a matter of degree (*Ibid.*, 266). Without arguing the point, I believe this idea needs clarification and that is the purpose of the present chapter and the following one.



have criticized parameter-resetting approaches to diachronic change in binding patterns on roughly the same grounds.<sup>2</sup> The objection is that such a parameter-setting model would make clear predictions with respect to how we could expect language change to occur and that those predictions are simply not borne out. Specifically, we would expect, on such an account, that whatever parameter was responsible for dictating whether a *self*-marked pronoun (in an A-position) required a local antecedent or not is set for each individual speaker as ‘+’ or ‘-’ (and not reset throughout the day). While this could still allow for a great deal of variation in a speech community, we would not expect systematic variation for individual speakers. But this exactly what we see; Keenan’s survey of late 15th, 16th, and early 17th century English authors illustrates this observation quite nicely.<sup>3</sup>

(5.1) *English self-marking frequencies for locally conjoint objects*

	<i>pro</i>	<i>pro+self</i>	%self
1495-1516 Skelton	57	99	63%
1533 Apologye	13	49	79%
1534 Berners	61	62	50%
1582 Learned	7	32	82%
1588-92 Marlowe	27	78	74%
1589-1605 Shakes	79	331	81%

Keenan’s own analysis of the shift in binding patterns and the genesis of reflexives in English involves reference to three “general forces of change”, *Inertia*, *Decay*, and *Pattern Generalization* as well as two “general semantic constraints on language” (Keenan, 2001, 1), *Constituency Interpretation* and *Antisynonymy*.

---

<sup>2</sup>For arguments for parameter-based language change, cf. the parameter-switch stories of Lightfoot (1989) and Platzack (1987), both of whom Levinson cites (2000, 362), or the proposals of Berwick (1985), Clark & Roberts (2003), Niyogi & Berwick (1997), and Briscoe (2000), whom Keenan (2001) mentions.

<sup>3</sup>The table in (5.1) is an abbreviated version of the one in (Keenan, 2001, 17). As Keenan (*Ibid.*, 18) duly points out, nothing about parameter-based models precludes predictions for scenarios like that shown in (5.1). Capturing those results would, however, seemingly require either an awfully large number parameters or something like the ‘noisy parameter settings’ framework of Yang (2000), which I will not discuss but which bears many pleasant similarities to a framework that I discuss at length below and ultimately rely upon.

(5.2) *Inertia*: Things stay the same.

*Decay*: Things wear out.

*Antisynonymy*: Different words mean different things.

*Constituency interpretation*: The constituents of an expression are semantically interpreted.

*Pattern Generalization*: A rule or paradigm that applies to a limited range of cases will extend to new ones.

Keenan's account of the local binding shift in English<sup>4</sup> runs roughly as follows: In the beginning (or at least around 750-1150 or so), *-self* was a contrast marker. It only marked contrast and always marked contrast. Through the influence of Decay, *pro+self* ceased to be obligatorily contrastive in A-positions and took on the meaning of ordinary pronouns in those positions.<sup>5</sup> Enter Antisynonymy. This force pushed toward a contrast in reference between the *pro+self* form and the bare pronoun.

Like Levinson's account, Keenan's account is intriguing because it attempts to explain binding patterns without making reference to Universal Grammar. Moreover, it offers an answer to at least one question that Levinson's account seemed to leave open, namely how to explain why forms like *myself* and *yourself* ever came to exist. Just as Levinson's M-principle and Blutner's weak bidirectional optimality cannot be responsible for such forms, Keenan's Antisynonymy cannot be the culprit either, since *me* and *myself* are indeed synonymous expressions – they both mean 'me'. Instead, Keenan attributes this step to Pattern Generalization; the pattern of consistent marking of third-person pronouns in bound A-positions generalized to pronouns of all person orientation.

However, although Keenan's Antisynonymy is pleasantly reminiscent of Grice's Avoid Ambiguity maxim, Levinson's M-principle and Blutner's weak

---

<sup>4</sup>I borrow the term "binding shift" from Keenan who uses the phrase to describe the roughly one hundred year period in the sixteenth century wherein we see a dramatic increase in the percentage of *self*-marked locally bound pronouns (from about 20% *self*-marked to around 80%) and a sharp decrease in occurrences of locally free ones.

<sup>5</sup>Keenan's Constituency Interpretation importantly prevents *pro+self* from losing its contrastive meaning in non-argument positions since, lacking any other meaning those positions, it would then mean nothing. In this way, the analysis of *pro+self* in A-positions and non-A-positions remains unified. Note that Constituency Interpretation is similar in some ways to the 'faithfulness' constraints seen in (Prince & Smolensky, 1993) and a great deal of other work in OT, though I will restrict my discussion to NPs in argument positions for the present purposes.

bidirectional optimality, Keenan never claims that the tendency toward non-synonymy and the interpretive differentiation between *pro* and *pro+self* that results from it is a byproduct of pragmatics, and since he does not invoke the DRP or anything like it, it is not immediately clear why such a “general force” or property is present in the first place. Keenan does, at one point, tentatively suggest (Keenan, 2001, 3) that Antisynonymy might be a byproduct of “gainful learning” whereby language users are inclined to learn a lexical inventory in a way that maximizes their expressive capabilities and thus in a way in which “new words mean new things”. However, I think that even if there is evidence for gainful learning, it would seem necessary to relate it to something like Horn’s division of pragmatic labor, Levinson’s M-principle, or Blutner’s notion of weak bidirectional optimality, so that it may offer specific predictions about the marked-form-for-marked-meaning pattern that shows up crosslinguistically and give an explicit explanation for why *self*-marked forms came to require local antecedents and bare pronouns came to resist them; Antisynonymy could have differentiated in either direction, one would think. Whatever the case, the actual mechanics of the attested differentiation between *pro* and *pro+self* need a formal explanation.

I think that the same can be said for Pattern Generalization. To claim that phenomena occur because of Pattern Generalization seems equivalent to saying that patterns generalize (for some unstated reason).

Antisynonymy and Pattern Generalization are *effects*, not causes.

In Chapter 6, I will sketch an account meant to address these issues. My aim is to exploit recent OT-based accounts of language learning and grammaticalization for the purpose of explaining how phenomena like the binding shift in English, where a Stage 1 language passes through Stage 2 and reaches Stage 3, may have occurred. In later discussion, I will suggest how the account might be further applied to issues which have surrounded the various binding analyses, including the issue of Keenan’s Pattern Generalization, the issue of multiple, discriminating reflexivization strategies like those seen in Dutch (cf. Chapter 2), as well as LDAs. With this, I hope to take steps toward explaining why certain binding patterns and certain trends in diachronic change are so common without reference to Universal Grammar and giving (at least a sketch of) a formal account of what the shift from Stage 1 to Stage 3 is and why it happens.

Both conceptually and formally, the account borrows heavily from earlier work of Zeevat & Jäger (2002), Zeevat (2002), Cable (2002), and Jäger (2003a), all of whom in one form or another advocate an OT-based evolu-

tionary picture of marking strategies that somehow permits for reference to something like the ‘stereotypes’ so often invoked in Levinson’s work. In particular, all of the aforementioned advocate the idea that *statistical bias* can play a role in the diachronic evolution of a grammar and the likelihood that certain types of marking patterns will show up. I will dedicate the remainder of the present chapter to introducing those ideas.

## 5.2 Optimization and Bias

Jäger (2003a, 7) points out that the idea that statistical frequencies could influence an evolutionary ‘choice’ in marking strategies might go back at least as far as Shannon (1948), who shows that an “optimal coding” in the information theoretic sense is one in which long codes are assigned to rare events and short codes to common ones. The same idea has been invoked by Zeevat & Jäger (2002) in an attempt to improve on the work of Aissen (1999, 2000),<sup>6</sup> who gives an account of differential case marking patterns in terms of the harmonic alignment of prominence scales.

A differential case marking (DCM) pattern is one in which the licensing of case marking is discriminatory in such a way that NPs with certain properties get case marked whereas NPs without those properties do not. The dimensions on which the discrimination is based vary crosslinguistically; the most common examples are person orientation, animacy, canonical role, and definiteness.

Examples include so-called ‘split-ergative’ systems like Dhargari (Austin, 1981), where animate objects are case marked but inanimate objects never are, or Dyirbal (Dixon, 1972), whose case marking system demands ergative case marking for non-local person subjects but never local person ones and accusative marking for local person objects but not for non-local ones.

Aissen claims that these patterns and others can be explained in terms of the alignment of multiple prominence scales. In a case where, say, animacy and canonical role were the relevant dimensions, prominence scales could be as follows.

### (5.3) *Prominence Scales*

#### a. Subject > Object

---

<sup>6</sup>(Aissen, 2000) was published as (Aissen, 2003).

b. Animate > Inanimate<sup>7</sup>

Aissen uses *harmonic alignment* of two scales like the ones in (5.3) to yield *harmony scales*, which represent the relative markedness of the various possible feature combinations. Technically the harmonic alignment function is defined as below.

(5.4) *Harmonic Alignment* (Prince & Smolensky, 1993, 136)

Given a binary dimension  $D_1$  with a scale  $X > Y$  on its elements  $\{X, Y\}$  and another dimension  $D_2$  with a scale  $a > b > c > \dots > z$  on its elements  $\{a, b, c, \dots, z\}$ , the *harmonic alignment* of  $D_1$  and  $D_2$  is the pair of harmony scales  $(H_x, H_y)$ , such that:

- a.  $H_x = X_a \succ X_b \succ X_c \succ \dots \succ X_z$  and
- b.  $H_y = Y_z \succ \dots \succ Y_c \succ Y_b \succ Y_a$ .

The harmonic alignment of (5.3a) and (5.3b) is thus:

(5.5) *Harmony Scales*

- Subject/Animate  $\succ$  Subject/Inanimate
- Object/Inanimate  $\succ$  Object/Animate

Aissen shows that the translation of harmony scales like the ones above into OT constraint subhierarchies can provide a means of capturing a universal tendency across languages to the effect that pair-types on the lower end of each harmony scale – i.e., *disharmonic* pair-types – are (a) relatively rare and (b) much more likely to be case marked compared to the *harmonic* pair-types at the high end of each scale.

(5.6) *Constraint Subhierarchies*

- \*Subject/Inan  $\gg$  \*Subject/Anim
- \*Object/Anim  $\gg$  \*Object/Inan

---

<sup>7</sup>I have oversimplified things here for the purpose of convenience, since the scale in (5.3b) is usually stated as ‘Human > Animate > Inanimate’. The more fine-grained distinction proves necessary sometimes for capturing patterns in languages like Yiddish (Aissen, 2000), where only human objects are case marked.

Following a line from classical markedness theory (Jakobson, 1939; Greenberg, 1966) to the effect that disharmonic feature combinations are generally either avoided *or* flagged with formal marking, Aissen posits two separate interpretations of the ‘\*X’-style constraints above that can serve to account for the general pattern which, by now, is quite familiar to us: marked meanings (i.e., unusual situations) get expressed by marked forms and unmarked meanings get expressed with unmarked ones.

(5.7) *Avoidance interpretation*

AVOID-Subject/Inan  $\gg$  AVOID-Subject/Anim  
AVOID-Object/Anim  $\gg$  AVOID-Object/Inan

(5.8) *Formal markedness interpretation*<sup>8</sup>

MARK-Subject/Inan  $\gg$  MARK-Subject/Anim  
MARK-Object/Anim  $\gg$  MARK-Object/Inan

It follows from (5.7) and (5.8) that things which are less likely to appear are things that are more likely to be marked.

Zeevat & Jäger (2002) have noted the relationship between Aissen’s work and the work of Blutner (2000b) and have proposed a functional explanation for some of the case marking patterns treated by Aissen based on the idea of bidirectional optimization and *bias*. The approach effectively derives the formal markedness interpretation of the constraint subhierarchies from the avoidance interpretation of those subhierarchies by (a) introducing a constraint into the interpretational evaluation procedure that can reflect a sensitivity to statistical states-of-affairs and (b) letting bidirectional optimization do its work in the usual way, yielding ‘weakly-optimal’ pairs in certain cases, which consist of marked forms and marked, i.e., rare, meanings.

The observation that marked combinations of features like Subject/Inan and Object/Anim are generally avoided cross-linguistically has been made in several places. As one example, Zeevat & Jäger (2002) and Jäger (2003a) consider the SAMTAL corpus of spoken Swedish (annotated by Östen Dahl), which exhibits the frequencies shown in (5.9) (per (Jäger, 2003a, 22)).

---

<sup>8</sup>Aissen actually formulates these constraints in terms of the *local conjunction* of two constraints. I discuss the local conjunction idea below; the difference is unimportant for the moment.

(5.9) *Clause-type frequencies in SAMTAL*

	<i>Anim/Obj</i>	<i>Inan/Obj</i>
<i>Anim/Obj</i>	300	17
<i>Inan/Obj</i>	2648	186

There is an obvious correlation between a pair's position on one of the harmony scales in (5.5) and the probability that it will show up in spoken Swedish and, it seems, any other language, though the strength of the correlation differs cross-linguistically.<sup>9</sup> With this in mind, Zeevat & Jäger (2002) and Zeevat (2002) basically take the avoidance interpretation of the '\*' for granted and take statistical asymmetries like the ones manifested in SAMTAL as universals of language use. Additionally, they propose an interpretational constraint that can serve to represent a linguistic sensitivity to those asymmetries. Zeevat (2002, 2) calls the constraint *Bias<sub>int</sub>*.

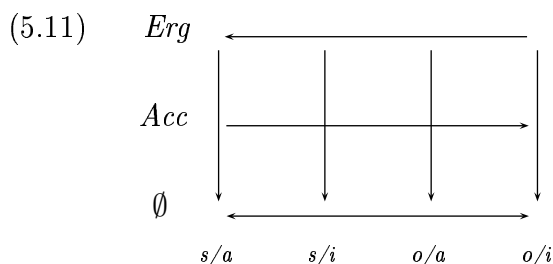
(5.10) *Bias<sub>int</sub>*: If a feature *f* is underspecified in a partially interpreted linguistic expression *L* in a context *c* then interpret *f*'s value as *v*, where *v* is the most probable value for *f*, given *L* and *c*.

In addition to *Bias<sub>int</sub>*, the Zeevat & Jäger (Z&J) account invokes two general, commonsense faithfulness and markedness constraints. Firstly, a constraint *Faith* favors faithful interpretations, e.g., ergative-marked NPs being interpreted as subjects (either animate or inanimate) and accusative-marked NPs being interpreted as (either animate or inanimate) objects. Secondly, a constraint like *\*Struct* will reflect a generative preference for unmarked outputs, regardless of the input.

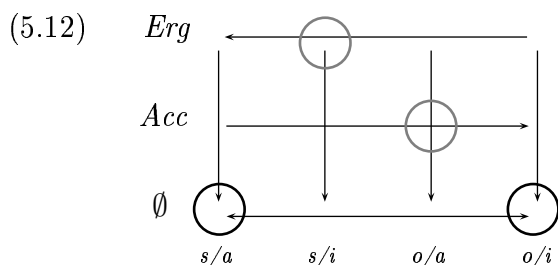
Due to the influence of *Bias<sub>int</sub>*, an individual living in a speech community where frequencies like those in SAMTAL were manifested would preferably interpret unmarked NPs as either animate subjects or inanimate objects. Restricting our attention to the set of inputs {*Erg*, *Acc*,  $\emptyset$ } (i.e., an ergative-marked NP, an accusative-marked NP, and an unmarked NP) and a set of outputs {*s/a*, *s/i*, *o/a*, *o/i*} (i.e., animate subjects, inanimate subjects, etc.), this gives us:

---

<sup>9</sup>Jacaltec (Craig, 1977) and Halkomelem (Gerds, 1988), for example, do not allow inanimate subject NPs to occur in transitive clauses at all. For further examples of evidence for the general avoidance of disharmonic feature combinations of this sort, cf. Fry (2001) or Lee (2001).



Following Z&J's commonsensical assumption that *Faith* dominates *Bias<sub>int</sub>*, we have four superoptimal pairs here, two of which are strongly bidirectionally optimal, viz.  $\langle \emptyset, \text{Subj/anim} \rangle$  and  $\langle \emptyset, \text{Obj/inan} \rangle$  and two of which are weakly bidirectionally optimal, viz.  $\langle \text{Erg}, \text{Subj/inan} \rangle$  and  $\langle \text{Acc}, \text{Obj/anim} \rangle$ .



In this way, the formal markedness interpretation of the Aissen-hierarchies is basically being derived via the complicity of *Bias<sub>int</sub>* and the mechanics of Bi-OT, rather than stipulated. For, just as it was not necessary to invoke any constraint like ‘MARK-*indirect-causation*’ to illustrate the kill/cause-to-die example, there is no need now to invoke a constraint like, say, MARK-Subj/Inan in order to predict a split ergative DCM pattern, since, in the Z&J picture, inanimate subjects will tend to get expressed in a marked way by virtue of bias and bidirectional optimization – they must be marked (under usual circumstances) because (under usual circumstances) unmarked forms would not be correctly interpreted as having the intended meaning (due to *Bias<sub>int</sub>*) and hence those forms would be blocked from expressing those meanings (due to bidirectional optimization).

On this view, though, marking would be triggered by blocking and thus we would not expect marking to occur in any cases where blocking does not. However, we have every reason to believe that, in certain contexts, overwhelming statistical bias from other sources – say, world knowledge – would override the general interpretational bias towards, say, interpreting



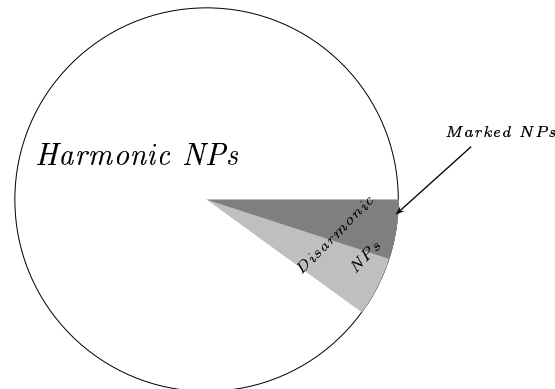
animate things as subjects, no matter how strong the more general biases got. (Consider a sentence like *A lightning bolt struck John*. We could assume that, regardless of the language, the description of lightning striking John would never require case marking to ensure a correct interpretation, even if word-order was not relevant.)

Zeevat & Jäger (2002, 11-12) have suggested that *Bias<sub>int</sub>* can offer a way of explaining the bridge between what Jäger (2003a, 10) calls ‘pragmatic’ DCM and ‘structural’ DCM, the former being a marking pattern wherein marking is employed only due to blocking and the latter being a pattern that is not restricted in this way. This is a problematic gap since, as Z&J themselves note (2002, 11) “Most case marking is obligatory”, i.e., structural.

Z&J envision a diachronic process whereby pragmatic DCM can become structural DCM due to the strengthening of bias(es) and a resulting “self-reinforcement” of the marking pattern in the following way:

Suppose one had a language wherein frequencies roughly like those of the SAMTAL corpus were manifested so that the harmonic combinations outnumbered disharmonic combinations nine-to-one. Suppose also that blocking actually did show up for 50% of the disharmonic combinations, but that blocking never showed up for harmonic combinations. From this and from the Z&J picture above, it would follow that a speaker of that language would mark 50% of disharmonic combinations and none of the harmonic ones.

(5.13)



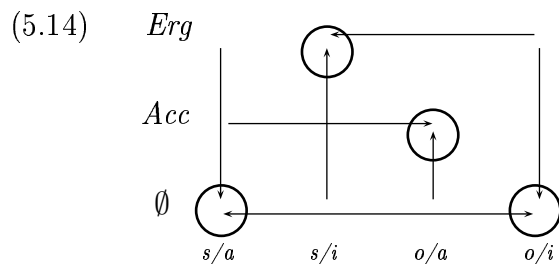
This 50% marking thus affects the statistical frequencies being observed by a language learner of the next generation, since the odds that an NP will be disharmonic is still 10%, but the odds that an *unmarked* NP is disharmonic is now only 5%. Ostensibly, the *Bias<sub>int</sub>* constraint of the next generation

would then represent an even stronger bias and this in turn would increase the chances of blocking and thus need for marking as well. In other words, the more marking there is in one generation, the more marking is necessary in the next generation. With each generation, there will be an increase in the percentage of marked disharmonic NPs until the marking becomes categorical. And, as Z&J put it:

“Once an optional marking strategy becomes non-exceptional and if it is functional, the marking makes itself more necessary and will normally become obligatory. It is then for the language learner at some point not distinguishable from a generation rule that requires marking certain combinations of features. As its original functional motivation and the process of self-reinforcement are not transparent to new language learners, learning it as a generation rule becomes the only option for new learners.”

–Zeevat & Jäger (2002, 11)

Once the ‘self-reinforcement’ of the marking strategy was complete, a nascent learner would learn a grammar wherein  $Bias_{int}$  militated unequivocally toward interpreting marked NPs as disharmonic and unmarked ones as harmonic. Furthermore, marking of disharmonic NPs would be obligatory, whereas marking harmonic NPs would be obligatorily avoided. We have a split ergative system:



However, the Z&J approach faces at least one or two major challenges.

Firstly, while it is reasonable to believe that biases do strengthen in the way that Z&J describe, there is actually no way to represent this in their system, since  $Bias_{int}$  is a single constraint and there is no way that constraint can be ranked to represent a strengthening of bias, nor any way that we could capture the fact that some biases might be stronger than others.

This problem is related to the more general criticism pointed to by Jäger (2003a, 10-11), namely that *Bias<sub>int</sub>* imports statistical sensitivity into the grammar by stipulation but does not provide an explanation of the mechanics of that sensitivity. Statistical bias is something that is *learned* from experience and there is no talk of how this is done, nor is it clear how the learning of that statistical knowledge cooperates with the learning of grammatical knowledge. As Jäger puts it:

“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... [F]requencies may help to explain why and how a certain grammar has been learned, but they are not part of this grammar.”

–Jäger (2003a, 11)

For these reasons, a more precise account of the grammaticalization of DCM systems using the idea of statistical bias has been sought. Cable (2002) was the first to seek it and he suggests an account of the shift from pragmatic-to-structural DCM based on the Gradual Learning Algorithm of Boersma (1998). Jäger (2003a) formalizes that account further and also formalizes the idea of interpretational bias itself by introducing an interpretational dimension to the story and showing how this dimension can play an important role in evolutionary learning and diachronic change. Both of these improvements on the original Z&J account involve specific reference to *stochastic OT*, a variation of OT discussed below.

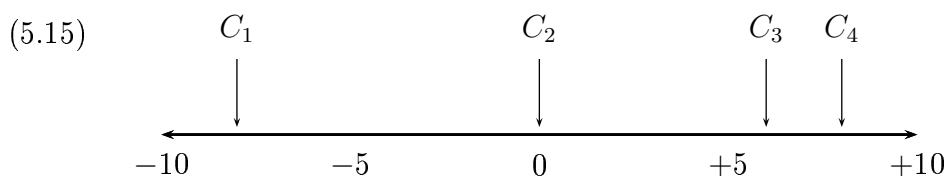
### 5.3 Stochastic OT

It is clear that whether we take pragmatic-to-structural DCM patterns or Stage 1-to-Stage 3 reflexive marking patterns as an example, the evolution of such patterns is a gradual process that, along the way, allows for a great amount of optionality and overlap with respect to the distribution of marked and unmarked forms. For this reason, formal accounts of those processes seem to require reference to some framework that would allow one to capture optionality and overlap of this kind. OT is equipped to deal with optionality in some ways, for two constraints can be unranked (write:  $C_1, C_2$ ) to produce optimality ties, whereby more than one candidate can be evaluated

as optimal, predicting free variation.<sup>10</sup> However, even this would not give us what we need to capture the kind of imbalanced optionality exhibited in the hypothetical case marking pattern we imagined above or the lopsided non-complementarity that existed in, say, OE *self*-marking or in the distribution of pronouns and R-expressions in a language like Thai (cf. Chapter 2). In those cases, the optionality does not amount to ‘free’ variation. Rather, one candidate is strongly preferred over the other, just not categorically so, and thus the variation is significantly constrained, not free. The *stochastic OT* of Boersma (1998) and Boersma & Hayes (2001) is a variation of standard OT which allows for empirical coverage for the kind of lopsided variation we are dealing with.

A stochastic OT grammar does not make a simple distinction between grammatical and ungrammatical expressions. Rather, it defines a probability distribution over a set of possible expressions and a particular expression is only technically ungrammatical if the grammar assigns that expression a probability of zero. Accordingly, an expression is preferred over another as a way of expressing a certain meaning just in case the probability for that expression is higher than that of its competitor, given the relevant meaning.

There are two major mechanical differences between stochastic OT and standard OT. Firstly, the ordinal ranking of standard OT is given up in stochastic OT and replaced by a *continuous ranking* of the relevant constraints, each one being assigned a real number called a *ranking value*. The various values of the various constraints not only serve to represent the hierarchical order of the constraints (higher values meaning higher ranks), but also to measure the distance between them.



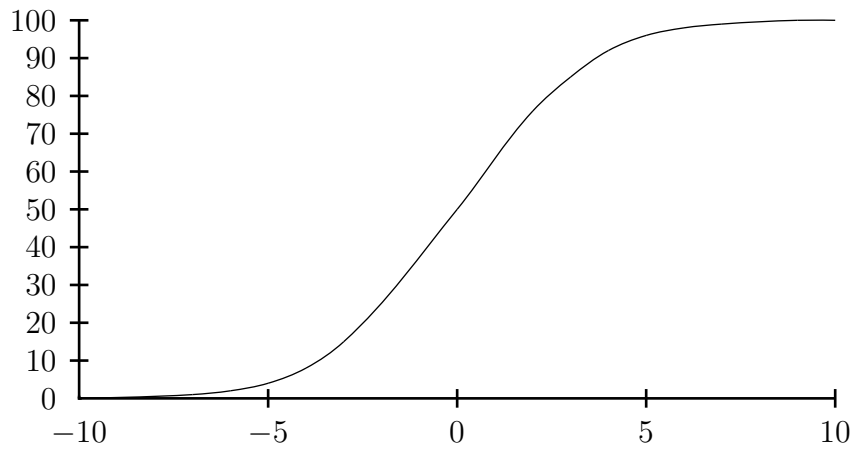
Secondly, stochastic OT employs *stochastic evaluation* such that, for each individual evaluation, the value of a constraint is modified with the addition of a normally distributed noise value. It is the strict hierarchical ranking of

---

<sup>10</sup>An alternative is *free ranking* (write:  $C_1 <> C_2$ ) whereby either  $C_1$  outranks  $C_2$  or vice versa (the choice being free), for any particular evaluation. As will be shown, stochastic OT essentially assumes free ranking for all constraints and probabilizes the ranking possibilities.

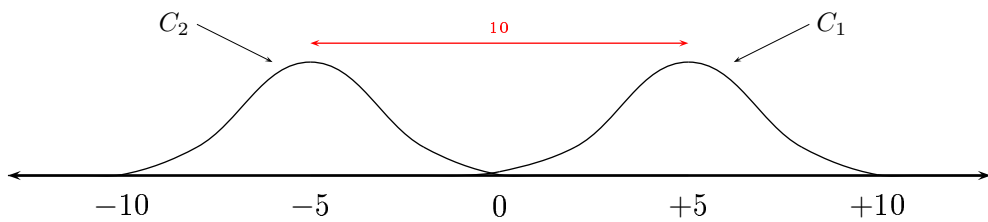
the constraints *after* adding the noise values that is responsible for the actual evaluation of the relevant candidates (for that individual evaluation). For any two constraints  $C_1$  and  $C_2$ , the actual probability that  $C_1$  will outrank  $C_2$  for any given evaluation is a function of the difference between their ranking values, where the dependency is the distribution function of a normal distribution such that  $\mu=0$  and  $\sigma=2\sqrt{2}$ , as is roughly depicted in (5.16).

(5.16)  $P(C_1 \gg C_2)$ , per  $C_1 - C_2$  (in %)



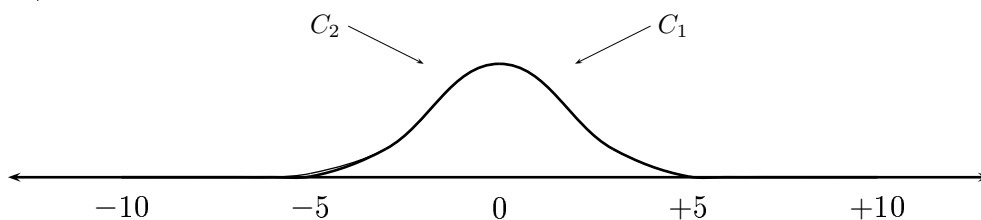
On this view, a categorical ranking for two constraints such that  $C_1 \gg C_2$  arises only when the ranking value of  $C_1$  is high enough compared to that of  $C_2$  that the probability of  $C_2$  outranking  $C_1$  for any given evaluation is virtually nil, say, 10 units.

(5.17)



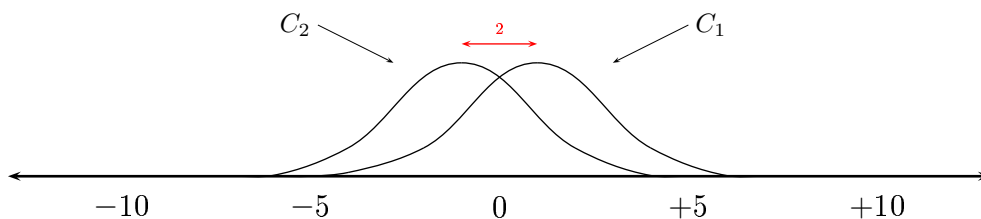
On the other hand, true free variation is predicted where two constraints have exactly the same ranking value.

(5.18)



Most importantly, however, are cases where the ranking values of two constraints are close enough to one another as to render the ranking of two constraints non-categorical, but where the ranking values are not equal either. In such cases, one predicts for optionality without predicting for totally free variation. For example, with the ranking schema in (5.19), below, we can expect about 76%-24% variation between candidates favored by  $C_1$  and those favored by  $C_2$ , since  $C_1$  outranks  $C_2$  by 2 units.

(5.19)



Aside from the advantages described above, stochastic OT has been shown to be compatible with a very powerful learning theory, the Gradual Learning Algorithm, due to Boersma (1998). This learning algorithm has in turn been invoked in explanations of grammaticalization in ways discussed below.

## 5.4 The GLA and Grammaticalization

Boersma's Gradual Learning Algorithm (GLA) is a method of systematically generating a stochastic OT grammar based on observed linguistic behavior and, thus, a theory of how a nascent learner could come to acquire knowledge of a grammar (i.e., knowledge of the ranking values of a set of constraints).

At any given stage of the learning process, the learner is assumed to have a hypothetical stochastic OT grammar in place. (By assumption, at

the beginning of the learning process the constraints are unranked, and thus equally strong.) Each time the algorithm is faced with the observation of some form-meaning pair, it uses the meaning as an input and generates some hypothetical output according to the hypothetical grammar currently in place. The algorithm then compares its hypothetical output to the actual output (i.e., the observed expression). If the hypothetical output and the observed expression are identical, no action is taken (for the hypothetical grammar is being ‘confirmed’ in such a case and does not need adjustment). However, if there is a ‘mismatch’ between the hypothetical output and the observed expression, the constraints of the learner’s grammar are adjusted in such a way that the observed output becomes more likely and the hypothetical output becomes less likely. In particular, all constraints that favor the observation are promoted by some small, predetermined amount, the *plasticity value*, and all those that favor the errant hypothesis are demoted by that amount. After a sufficient number of inputs, the learned grammar will converge into one that assigns (roughly) the same probabilities to all the same candidates as the grammar which generated the representative sample that served as the learning data for the learned grammar. The learned grammar is thus a (perhaps imperfect) replica of the grammar that generated the learning corpus.<sup>11</sup> A grammar can be said to have *converged* just in case further observations no longer induce significant adjustments of the learner’s hypothetical grammar.

Cable (2002) proposes to explain the shift from pragmatic-to-structural DCM through an evolutionary story which combines the GLA with the ideas of bidirectional optimization and bias. The story assumes that a learner has access to five generative constraints like those proposed by Aissen (2000), stated again in a simplified form below.

(5.20) \**Struct*: Avoid morphological structure.

MARK-Subject/Anim: Case mark animate subjects.

MARK-Subject/Inan: Case mark inanimate subjects.

MARK-Object/Anim: Case mark animate objects.

MARK-Object/Inan: Case mark inanimate objects.

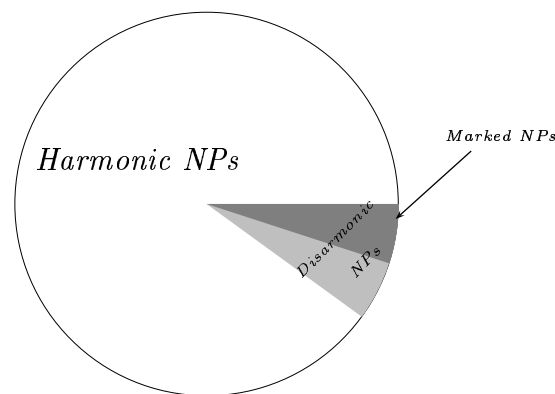
Cable imagines a language wherein pragmatic DCM had reached its maximum but where structural DCM was totally absent and thus case specifica-

---

<sup>11</sup>Typically, it is assumed that the learner’s grammar and his ‘teacher’s’ grammar consist of the same set of constraints.

tion is always marked when an unmarked NP is not optimally recoverable due to  $Bias_{int}$ , but never marked otherwise.<sup>12</sup> If we take such a marking strategy for granted along with SAMTAL-like frequencies with respect to the distribution of the various NP-types and if we assume again that, in the hypothetical language, harmonic combinations could always be expressed with unmarked NPs and still be optimally recoverable but that disharmonic combinations were only recoverable, say, 50% of the time, we have a scenario just like the hypothetical Z&J picture discussed above.

(5.21)



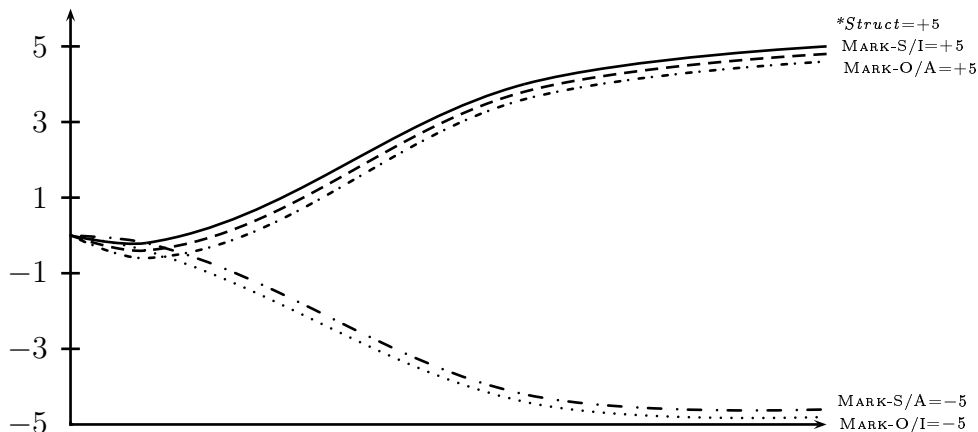
Supposing that the learner is exposed to learning data where such frequencies were present, the learner could learn the aversion to case marking harmonic NPs as a categorical preference so that *\*Struct* outranked both MARK-Subject/Anim and MARK-Object/Inan by 10 units or so. On the other hand, the constraints which pertain to disharmonic NPs, MARK-Subject/Inan and MARK-Object/Anim, would be learned as having a ranking value comparable to that of *\*Struct*.

---

<sup>12</sup>Cable's example involves the local/non-local person dimension, not the animacy dimension, though the difference is immaterial to the argument.

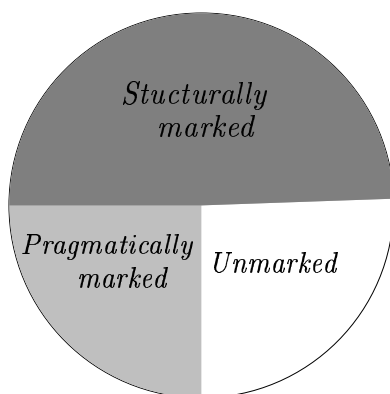


(5.22) *Hypothetical learning curves (first generation)*



This learner's grammar would lead him to mark 50% of disharmonic NPs regardless of whether blocking occurs or not. Moreover, he would also mark all of the disharmonic NPs where blocking does show up, by virtue of bidirectional optimization. By assumption, 50% of disharmonic NPs are ambiguous and, thus, 75% of disharmonic NPs will get case marked in the learner's speech; 50% will be 'structurally' marked and 25% will be 'pragmatically' marked.

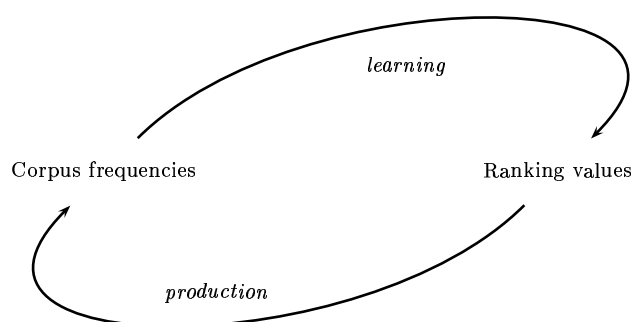
(5.23) *Marked versus unmarked disharmonic NPs (first generation)*



As is implicit in the Z&J story and explicit in the later the work of Jäger (2003a), Cable effectively assumes the *Iterated Learning Model* (ILM)

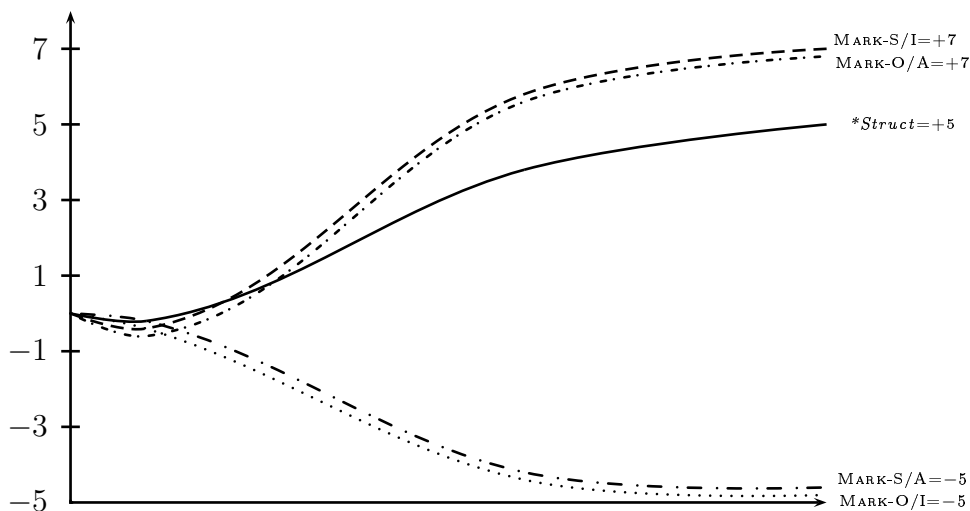
of language evolution due to Kirby & Hurford (1997). That model takes each generation of learners to be one turn in a cycle of language evolution and, by applying a learning algorithm to the output of one cycle, one may produce a second cycle, and then a third, a fourth, and so on.

(5.24) *Iterated Learning Model* (Kirby & Hurford, 1997)



A second-generation learner who was exposed to frequencies per (5.23) will learn a grammar that reflects the 75% marking of disharmonic NPs, not 50%.

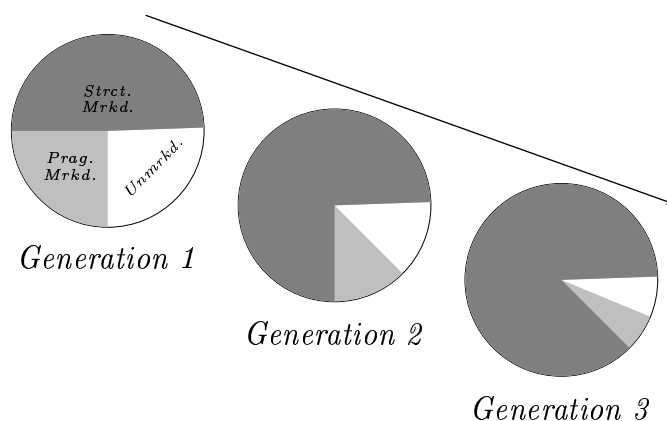
(5.25) *Hypothetical learning curves (second generation)*



The correlation between blocking and marking is again lost, since the effects of bidirectional optimization have been learned just as if they were

the effects of unidirectional optimization. Moreover, this second generation speaker will also employ pragmatic DCM on top of structural DCM, and thus he will case mark 87.5% of all disharmonic NPs rather than 75% like the previous generation did. An iteration of this cycle will see an increase in marked disharmonic NPs with each generation.

(5.26) *Pragmatic-to-structural DCM*



Cable’s approach describes a precise mechanism of grammaticalization and this is something that we have seen to be lacking in both the earlier accounts of DCM per Zeevat & Jäger (2002) and Zeevat (2002) as well as in Levinson’s (2000) and Keenan’s (2001) discussions of the genesis and grammaticalization of reflexives.

However, the issues surrounding the constraint  $Bias_{int}$  still leave serious questions open. The constraint  $Bias_{int}$  is not shown in the hypothetical learned grammars above and it would make no sense to include it, since  $Bias_{int}$  is an interpretational constraint and, in a GLA-based picture like the one above, a learner is not learning interpretational preferences, he is only learning generative ones. The actual learning of the biases themselves and their appearance in the grammar is still being taken for granted. Without a formal way of integrating the idea of bias into the grammar, there is still no way to explain why, say, Subj/Inan pairs are really identified as disharmonic things since, without some kind of interpretational learning, the learner has no way of forming or recording interpretational preferences in a way that can be expressed via OT constraint rankings.

Jäger (2003a, et al.) has proposed a way to deal with this issue. Specifically, he proposes an account of grammaticalization based on a formally

spelled out and formally implemented *bidirectional* version of the GLA wherein the same the five constraints proposed by Aissen can – again per the assumptions of Wilson (2001) and Buchwald et al. (2002) – function as generative *and* interpretational constraints. A learner can then learn statistical biases and represent those biases through the ranking of those constraints.

## 5.5 BiGLA

Jäger’s account of the shift from pragmatic-to-structural DCM employs a bidirectional version of the GLA, called the Bidirectional Gradual Learning Algorithm, or BiGLA. It is an attempt to overcome the issues surrounding the idea of interpretational bias by extending the notion of bidirectional optimality to the learning process in two separate ways.

Firstly, just as before, the notion of *bidirectional evaluation* is imported into the learning algorithm by stipulating a recoverability restriction for optimality. Forms are disqualified as candidates when they are not optimally recoverable as the intended meaning and at least one other form is. The bidirectional optimization in this case is *asymmetric* in the sense that there is no analogous blocking mechanism for meaning candidates.<sup>13</sup> Officially this is stated as below.

(5.27) *Asymmetric bidirectional optimality* (Jäger, 2003a, 19)

- a. A form-meaning pair  $\langle f, m \rangle$  is *hearer optimal* iff there is no pair  $\langle f, m' \rangle$  such that  $\langle f, m' \rangle \succ \langle f, m \rangle$ .
- b. A form-meaning pair  $\langle f, m \rangle$  is *optimal* iff either  $\langle f, m \rangle$  is hearer optimal and there is no distinct pair  $\langle f', m \rangle$  such that  $\langle f', m \rangle \succ \langle f, m \rangle$

---

<sup>13</sup>Asymmetric versions of bidirectional OT may have started with Wilson (2001). The asymmetric variation of bidirectionality allows one to avoid a few puzzles faced by the symmetric version. One such puzzle is the so-called ‘*Rat/Rad* problem’: under the symmetric picture, we expect the correct interpretation of an utterance like German *Rad* (‘wheel’) (homophonous with *Rat* (‘council’)) to be blocked, since the pronunciation of *Rad* as /rat/ presumably violates a faithfulness constraint which the pronunciation of *Rat* does not. Beaver & Lee (2003) point out that one consequence of asymmetric OT is the loss of empirical coverage for partial blocking of the kind that was captured by Levinson’s M-principle or Blutner’s weak optimality, where the marked forms get interpreted as marked meanings since the unmarked meanings are blocked. However, the asymmetric picture does not preclude an *evolutionary* explanation of marked-forms-for-marked-meanings pattern, and this is more the aim of the Z&J, Zeevat, Cable, and Jäger stories anyway.

and  $\langle f', m \rangle$  is hearer optimal, or no pair is hearer optimal and there is no distinct pair  $\langle f', m \rangle$  such that  $\langle f', m \rangle \succ \langle f, m \rangle$ .

Secondly, learning in the BiGLA is *bidirectional learning* in the sense that a learner not only evaluates candidate forms with respect a hypothetical grammar, but also candidate meanings. For this reason, where a learner is faced with a learning datum,  $\langle f, m \rangle$ , he now not only compares the actual form,  $f$ , with some hypothetical output,  $f'$ , produced by his hypothetical grammar, but also produces a hypothetical meaning,  $m'$ , and compares it to the actual observed meaning,  $m$ .<sup>14</sup> Learning effects may take place that involve the adjustment of constraints that evaluate meanings in addition to those which evaluate forms, and, crucially, some constraints may be affected by both hearer- and speaker-learning modes. Jäger's BiGLA learning algorithm can be represented schematically as the six-stage procedure below.

(5.28) *BiGLA* (Jäger, 2003a, 20-21)

1. *Initial state*

All constraint values are set to 0.

2. *Step 1: Observation*

The algorithm is presented with a learning datum, a fully specified input-output pair  $\langle f, m \rangle$ .

3. *Step 2: Generation*

For each constraint, a noise value is drawn from a normal distribution  $N$  and added to its current ranking. This yields a *selection point*. Constraints are ranked by descending order of the selection points. This yields a linear order of the constraints  $C_1 \gg \dots \gg C_n$ . Based on this constraint ranking, the grammar generates a hypothetical output,  $f'$ , for the observed input  $m$  and a hypothetical output,  $m'$ , for the observed input  $f$ .

4. *Step 3: Comparison*

---

<sup>14</sup>An important assumption is required here, namely that the learner will somehow successfully determine correct meaning of the observed form. Interpretational learning would not be possible if we could not assume that this happened at least some of the time. Cases where the observed meaning is not successfully recovered are ignored for the present purposes.

If  $f' = f$ , nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum  $\langle f, m \rangle$  with the hypothetical pair  $\langle f', m \rangle$ .

If  $m' = m$ , nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum  $\langle f, m \rangle$  with the hypothetical pair  $\langle f, m' \rangle$ .

5. *Step 4: Adjustment*

All constraints that favor  $\langle f, m \rangle$  over  $\langle f', m \rangle$  are increased by the plasticity value. All constraints that favor  $\langle f', m \rangle$  are decreased by the plasticity value.

All constraints that favor  $\langle f, m \rangle$  over  $\langle f, m' \rangle$  are increased by the plasticity value. All constraints that favor  $\langle f, m' \rangle$  are decreased by the plasticity value.

6. *Final state*

Steps 1-4 are repeated until the constraint values stabilize.

Jäger shows how an explicit combination of the BiGLA and the ILM can be applied to give a formal account of the shift from pragmatic-to-structural DCM that avoids the conceptual problems of Zeevat's *Bias<sub>int</sub>* constraint. The idea is to let the generative constraints proposed by Aissen serve as interpretational constraints as well as generative constraints, whereupon they will then be subject to adjustment in hearer-mode learning as well as speaker-mode learning. To do this, he follows Aissen's original idea of stating constraints of the form 'MARK- $X$ ' as the *local conjunction* of two constraints  $*X$  and  $*\emptyset$ , the former militating against  $X$  (whatever that is) and the latter penalizing the absence of case specification.<sup>15</sup>

(5.30)  $*_{s/a, \emptyset}$ : NPs denoting animate subjects are case marked.

$*_{s/i, \emptyset}$ : NPs denoting inanimate subjects are case marked.

---

<sup>15</sup>Aissen credits the original idea of stating the constraints this way to Paul Smolensky, who proposed the idea of local constraint conjunction in Smolensky (1995):

(5.29) *Local constraint conjunction*

The local conjunction of  $C_1$  and  $C_2$  in domain  $D$ ,  $C_1 \&_D C_2$ , is violated when there is some domain of type  $D$  in which  $C_1$  and  $C_2$  are violated.

For criticisms of local constraint conjunction, cf. Zeevat & Jäger (2002) or Cable (2002), though stating constraints this way can still be done for convenience without any harm.

- \**o/a*, $\emptyset$ : NPs denoting animate objects are case marked
- \**o/i*, $\emptyset$ : NPs denoting inanimate objects are case marked.

The constraints above are fairly self-explanatory. With respect to generative optimization, each constraint of the form  $*x,\emptyset$  militates against the output of an unmarked NP, given some input  $x$ . With respect to interpretational optimization, each constraint of the form  $*x,\emptyset$  militates against interpreting an unmarked NP as  $x$ .

Hearer-mode learning will now be able to register interpretational biases with respect to unmarked NPs by ranking the constraints appropriately amongst one another.

On the other hand, speaker-mode learning will affect the ranking of the four constraints in (5.30) in relation to  $*Struct$  in a way that reflects the generative preferences.

We can assume that we are dealing with a scenario like the one discussed by Zeevat & Jäger (2002) and Cable (2002) wherein 50% of disharmonic NPs are case marked, but harmonic NPs never are. If we adopt SAMTAL-like frequencies, per (5.9), this would give us a corpus in which the absolute numbers would look as below.

(5.31) *Hypothetical training corpus (based on SAMTAL)*

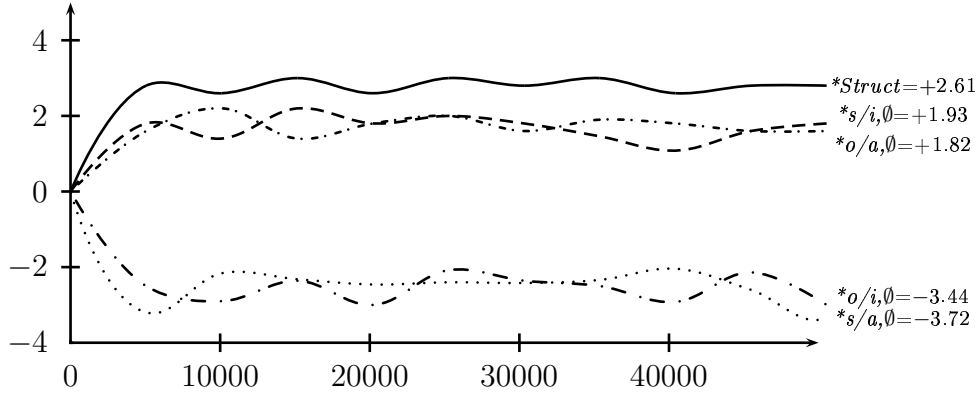
	<i>Erg/Acc</i>	<i>Erg/<math>\emptyset</math></i>	$\emptyset/Acc$	$\emptyset/\emptyset$
<i>Subj/Anim-Obj/Anim</i>	0	0	150	150
<i>Subj/Anim-Obj/Inan</i>	0	0	0	2648
<i>Subj/Inan-Obj/Anim</i>	4	4	4	3
<i>Subj/Inan-Obj/Inan</i>	0	93	0	93

Feeding BiGLA with sixty-thousand inputs drawn at random based on the frequencies in (5.31) yielded the learning curves below.<sup>16</sup>

---

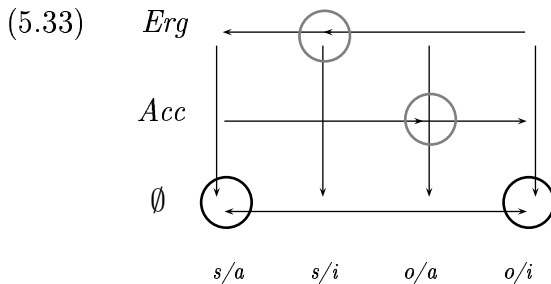
<sup>16</sup>The simulation – and all the simulations in this dissertation – are conducted using *evolOT*, which is an implementation of the (Bi)GLA developed by Gerhard Jäger. Currently, the software is available for download at no cost from <http://www.ling.uni-potsdam.de/~jaeger/evolOT/>.

(5.32) *Bidirectional learning curves (based on (5.31))*



Consider the two dimensions of optimization given a grammar like the one in (5.32). On the interpretational side, hearer-mode learning has resulted in the constraints  $*s/i, \emptyset$  and  $*o/a, \emptyset$  being ranked about evenly, and both are ranked significantly higher than  $*s/a, \emptyset$  and  $*o/i, \emptyset$ . This reflects a very strong preference for interpreting unmarked NPs as either animate subjects or inanimate objects, i.e., interpreting unmarked NPs as harmonic NPs.

On the generative side,  $*Struct$  greatly outranks  $*s/a, \emptyset$  and  $*o/i, \emptyset$ , meaning that harmonic NPs will not get marked. But  $*Struct$  is ranked only slightly above  $*s/i, \emptyset$  and  $*o/a, \emptyset$ , meaning that structural DCM will be employed for disharmonic NPs about 60% of the time. We have:<sup>17</sup>



A set of simulated output frequencies based on the grammar in (5.32) looked as below.

<sup>17</sup>For convenience, I have left *Faith* out of this experiment, assuming it is never violated.



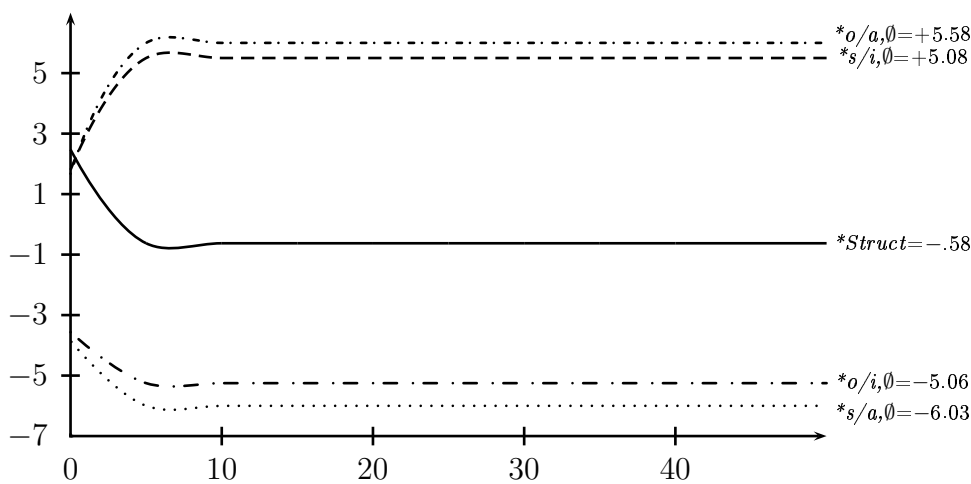
(5.34) *Frequencies of (5.32)*

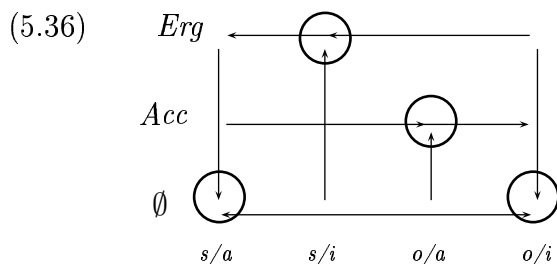
	<i>Erg/Acc</i>	<i>Erg/∅</i>	<i>∅/Acc</i>	<i>∅/∅</i>
<i>Subj/Anim-Obj/Anim</i>	10	0	284	6
<i>Subj/Anim-Obj/Inan</i>	11	81	143	2413
<i>Subj/Inan-Obj/Anim</i>	13	2	0	0
<i>Subj/Inan-Obj/Inan</i>	13	170	0	3

Note that though structural DCM is warranted by the grammar only about 60% of the time, almost all disharmonic NPs are marked here. Thus, we know that pragmatic DCM is responsible for the rest of the marking. This rapid grammaticalization is due to the fact that our experiment reflects no distinction between unmarked disharmonic NPs which are blocked due to the fact that they are not contextually recoverable from those which are not blocked. Rather, in the experiment above, all disharmonic unmarked NPs are assumed to be blocked.

It is enough to illustrate the point, though, of how using the frequencies in (5.31) as an *Ur*-corpus and applying the BiGLA and the ILM can yield an evolved grammar with obligatory, structural marking, like the one in (5.35), below.

(5.35) *Evolution (50 generations)*





Again, the evolution of pragmatic DCM into structural DCM in this case took place almost immediately because, as noted, we have made no distinction between unmarked disharmonic NPs which are blocked and those which are not. Integrating fine-grained contextual constraints into the picture must remain the area of further research and I do not want to dwell on it here, since an explanation of differential case marking is not my ultimate goal.

I will note however, that it is very common for a grammar like the one above to evolve into one in which *both* disharmonic *and* harmonic NPs are obligatorily case marked, and this is not representative of a crosslinguistically typical case marking strategy and is not in line with the marked-form-for-marked-meaning pattern that we might expect (or at least want) to see.

Such an issue becomes much even more acute when we consider the emergence of case markers themselves and imagine a situation (as Jäger (2003a, 34-38) himself discusses) in which a language is not endowed with both ergative and accusative morphemes, but rather has one marker which is not yet lexically specified as a case marker. (We could compare this to something like a ‘Stage 1’ system in the Carden & Stewart (1988)/Levinson (2000) senses, discussed above.)

Imagining a simplified version of an experiment conducted by Jäger (2003a), we can suppose we had a corpus with SAMTAL-like asymmetries, but in a language that possessed only one marking morpheme that had not yet taken on any bona fide case specification. We can suppose (per Jäger) for the purpose of experimentation that the distribution of the marking was *non-differential*, i.e., disharmonic and harmonic NPs are marked equally, say 50% of the time. We might then expect or at least hope that evolutionary iterated BiGLA-learning applied to such a training corpus would, after a sufficient number of generations, evolve into an unequivocal marked-form-for-marked-meaning pattern. We would then have before us a potential explanation for that pattern, as it could be taken to be a result of bidirectional learning

and the bias effects that go along with it, plus the effects of bidirectional optimization.

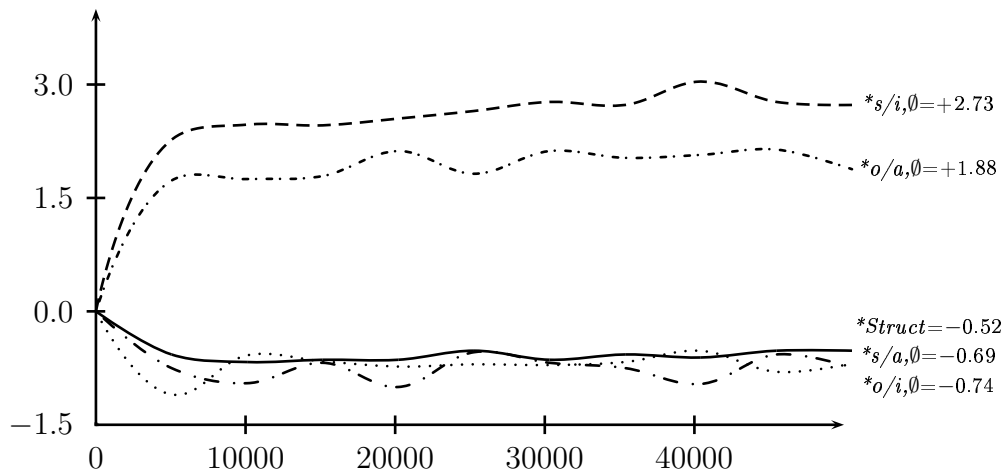
Let a hypothetical corpus have the absolute numbers in (5.37).

(5.37) *Training corpus with non-differential case marking*

	$M/M$	$M/\emptyset$	$\emptyset/M$	$\emptyset/\emptyset$
<i>Subj/Anim-Obj/Anim</i>	300	300	300	300
<i>Subj/Anim-Obj/Inan</i>	2648	2648	2648	2648
<i>Subj/Inan-Obj/Anim</i>	17	17	17	17
<i>Subj/Inan-Obj/Inan</i>	186	186	186	186

The letting BiGLA do its work with (5.37) yielded the learning curves below.

(5.38) *Bidirectional learning, per (5.37)*



Note that preference for marking disharmonic pairs is, again, immediate and is almost categorical:

(5.39) *Frequencies, per (5.38)*

	$M/M$	$M/\emptyset$	$\emptyset/M$	$\emptyset/\emptyset$
<i>Subj/Anim-Obj/Anim</i>	654	6	510	30
<i>Subj/Anim-Obj/Inan</i>	3846	2065	1916	2765
<i>Subj/Inan-Obj/Anim</i>	66	1	1	0
<i>Subj/Inan-Obj/Inan</i>	424	317	1	2

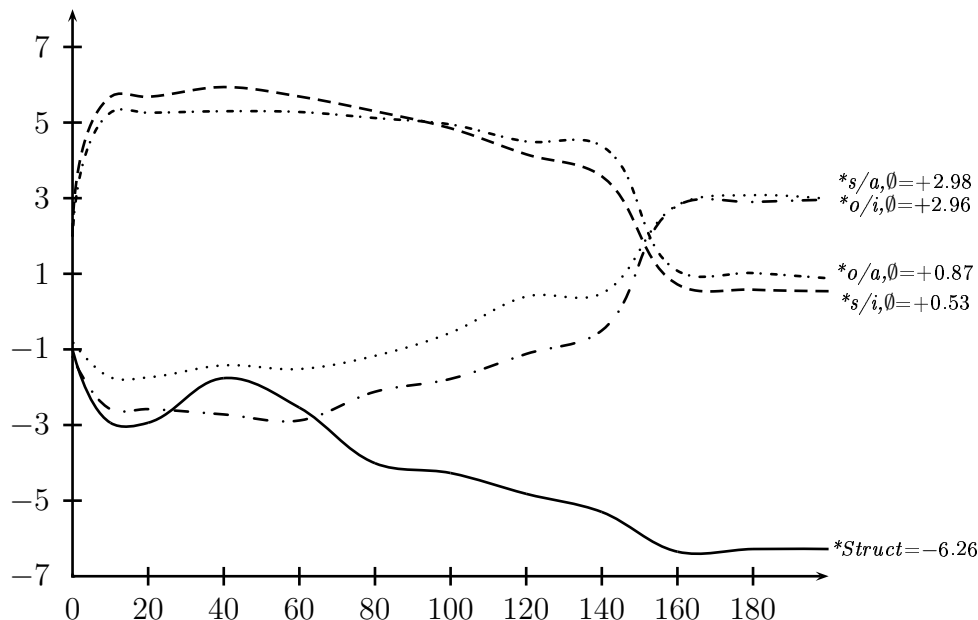
The marked meanings now demand marked forms.

The reason for this pattern coming to be is that the corpus in (5.37) is essentially an impossible ‘puzzle’ for BiGLA to solve, given the constraints it has to work with in this case. On the one hand, speaker-mode learning will prefer to keep all the constraints at zero; since the case marking is indiscriminately 50%-50% for all NPs, the five constraints should all be equally ranked in order to ‘satisfy’ speaker-mode. On the other hand, hearer-mode learning will be ‘separating’ two pairs of constraints – viz. the pair ( $*_s/i, \emptyset$ ,  $*_o/a, \emptyset$ ) and the pair ( $*_s/a, \emptyset$ ,  $*_o/i, \emptyset$ ) – in order to reflect the interpretational bias that will be learned from the asymmetries in the training corpus frequencies. This grammar cannot converge, then, because speaker-mode and hearer-mode are in conflict with one another.

I will elaborate more on this last point in Chapter 6, since I believe that the interaction of markedness constraints and constraints that represent interpretational bias, and the dissonance this interaction creates in a bidirectional learning framework, can facilitate an account of Horn’s ‘division of pragmatic labor’ (Horn, 1984), i.e., the marked-forms-for-marked-meanings pattern, and that this can provide a promising way of improving on Levinson’s pragmatic account of binding patterns. I must point out now though, that in my experience with Aissen’s constraints and BiGLA experiments, the intuitive evolutionary outcome – i.e., that marked forms will come to pair with marked meanings – is never the permanent one. Rather, on top of the fact that the learned grammar based on (5.37) has developed pragmatic marking at a somewhat unintuitively fast pace, the evolutionary path of such a grammar also virtually never leads to split ergativity. Rather, while the disharmonic inputs will indeed continue to pair with marked forms, the harmonic meanings eventually will too.

Just as an example, using (5.37) as a training corpus for the original generation and executing two hundred generations of iterated learning per Jäger’s evolutionary OT yielded the following.

(5.40) *Evolution of (5.38)*



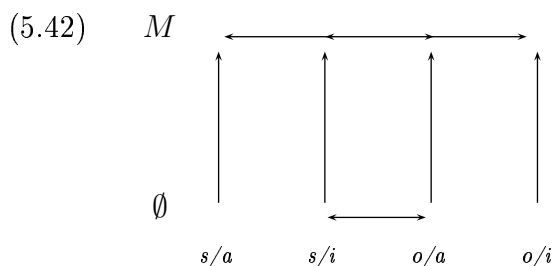
This grammar is one in which (almost) *all* NPs are marked with the underspecified morphological marker we are imagining.

(5.41) *Frequencies, per (5.40)*

	$M/M$	$M/\emptyset$	$\emptyset/M$	$\emptyset/\emptyset$
<i>Subj/Anim-Obj/Anim</i>	1196	4	0	0
<i>Subj/Anim-Obj/Inan</i>	10589	2	1	0
<i>Subj/Inan-Obj/Anim</i>	68	0	0	0
<i>Subj/Inan-Obj/Inan</i>	740	0	4	0

Moreover, the grammar has stabilized in a rather unintuitive way, for note that, per the rankings, marking harmonic feature combinations is demanded even more vigorously than marking disharmonic ones. The Aissen-style hierarchies have not been preserved.

This gives us a scenario that actually involves four bidirectionally optimal pairs, though certainly not in the way we expected.



This is certainly not the intuitive outcome, since the marker in this hypothetical grammar is, in effect, ‘wasted’, as it is a structural liability on the speaker which reaps no benefit for the hearer. Thus, this picture, as it stands, seems unsuitable for describing the evolution of split ergative systems or ‘split marking’ strategies in general and would hence be equally unsuitable for capturing the type of pattern that is the ultimate goal here, namely reflexive marking patterns.

I believe that the heart of the problem lies with the constraints themselves. If we use only Aissen’s repertoire of constraints to represent hearer-bias then the learner has no constraint telling him how to interpret marked NPs. Jäger (2003a, 34), for his part, suggests two additional constraints:<sup>18</sup>

(5.43)  $M \Rightarrow subj$ : Marked forms are subjects.

$M \Rightarrow obj$ : Marked forms are objects.

But even this would not give the learner any bias potential with respect to whether to prefer to interpret marked forms as animate or inanimate, and presumably he should be able to show bias in this respect as well.

I believe the most direct way to address the issue is to simply assume that, in addition to the single markedness constraint *\*Struct*, a learner is equipped with a totally comprehensive, totally neutral set of *bias constraints*, which will allow him to accurately ‘record’ biases about whatever properties he is sophisticated enough to recognize.

If we restrict our attention to only the properties under discussion, i.e., canonical role and animacy, then we need only Aissen’s four constraints above plus the following four to achieve a genuine ‘bias counter’.

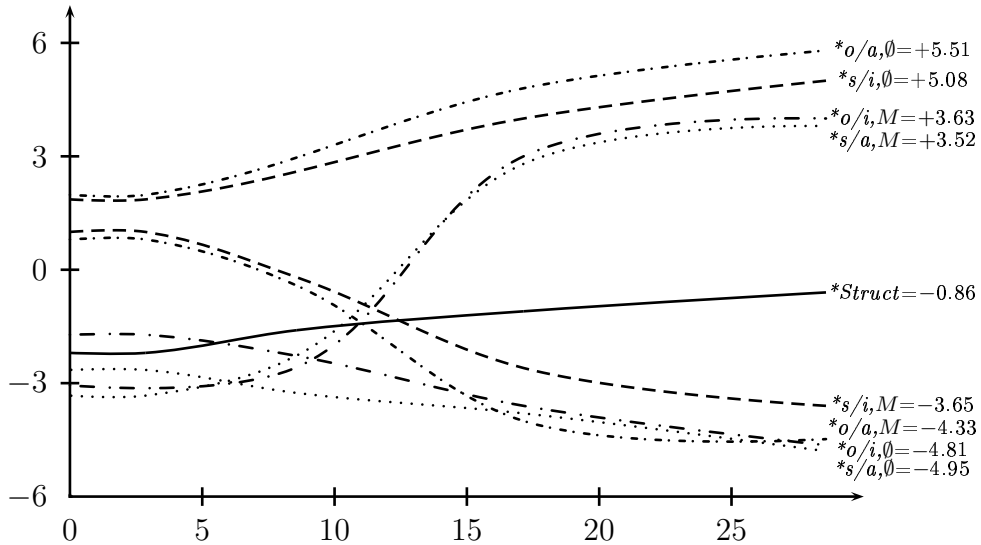
<sup>18</sup>Actually, Jäger suggests four additional constraints, two of which are his ‘*SO*’ and ‘*OS*’, which are constraints related to word order. I am leaving word order considerations out of the present discussion for simplicity, and thus simply assuming that, in our hypothetical language, it is irrelevant.

- (5.44)  $*_{s/a,M}$ : NPs denoting animate subjects are not case marked.  
 $*_{s/i,M}$ : NPs denoting inanimate subjects are not case marked.  
 $*_{o/a,M}$ : NPs denoting animate objects are not case marked.  
 $*_{o/i,M}$ : NPs denoting inanimate objects are not case marked.

Letting these four constraints compete directly with the constraints  $*_{s/i,\emptyset}$ ,  $*_{o/a,\emptyset}$ , and so on is, I think, a way of doing things that is at least much more in the spirit of Zeevat's *Bias<sub>int</sub>* constraint in that it provides a completely inclusive reservoir of constraints that will simply reflect the statistical asymmetries by way of their rankings relative to one another. Moreover, I think that the benefit of this approach will carry over into the application of evolutionary BiGLA to the area of binding phenomena in ways that I will elaborate on in the next chapter.

As for the actual outcome of an experiment that included the usual constraints, plus those in (5.44), using the frequencies in (5.37) as an ancestor corpus and executing thirty generations of iterated learning yielded results that look roughly as below.

(5.45) *Evolution, per (5.37) (30 generations)*



Now *this* is a split-ergative system! Disharmonic feature combinations will warrant marked forms and harmonic feature combinations will not. Marked forms for (and only for) marked meanings. Moreover, the grammar shown in (5.45) showed little or no variation once it had stabilized, even after thousands of generations.

(5.46) *Frequencies (per (5.45))*

	<i>M/M</i>	<i>M/∅</i>	<i>∅/M</i>	<i>∅/∅</i>
<i>Subj/Anim-Obj/Anim</i>	0	0	1200	0
<i>Subj/Anim-Obj/Inan</i>	0	0	0	10592
<i>Subj/Inan-Obj/Anim</i>	68	0	0	0
<i>Subj/Inan-Obj/Inan</i>	0	744	0	0

As I noted before, I think that the strategy employed to get a result like this can help with telling a story about the genesis and grammaticalization of reflexive marking patterns based on evolutionary bidirectional learning. I turn to this story in the following chapter.



# Chapter 6

## Bias, Bidirectionality, & Binding Phenomena

### 6.1 Introduction

My main aim below is to show how, instead of case marking strategies, we can tell a grammaticalization story involving Jäger's (2003a) BiGLA-based evolutionary OT and the ideas of Zeevat & Jäger (2002), Zeevat (2002), and Cable (2002) about the role of statistical bias in grammar to address questions surrounding the evolution of reflexive marking strategies.

I will focus primarily on one experiment meant to simulate the transition from an optional and infrequent marking strategy like that of Old English or some other 'Stage 1' language into a pattern of obligatory structural marking like the one attested in Modern Standard English or some other 'Stage 3' language, in the sense of Carden & Stewart (1988) and Levinson (2000). Though excessively simple, I think that the experiment illustrates in a clear way how a comprehensive reservoir of bias constraints can conspire with one or more markedness constraints within a framework of evolutionary bidirectional learning to predict the marked-form-for-marked-meaning pattern evidenced in reflexive-marking strategies, case marking strategies, and so many other facets of language use.

It might be a pleasant thing if my optimism is justified, for, as I have discussed above, though the marked-things-for-marked-things pattern has often been explained in terms of pragmatics – e.g., Horn's division of pragmatic labor, Levinson's M-principle, or Blutner's weak bidirectional optimality, all

discussed above – a precise explanation of exactly how it could manifest itself in (the evolution of a) grammar has been somewhat elusive.

Before concluding, I will suggest how the account might be extendable to other aspects of attested patterns of binding behavior such as Keenan’s ‘Pattern Generalization’, whereby a marking pattern that applies to a limited range of cases extends to new ones, the pattern of discriminatory SE/SELF distribution in a language like Dutch, discussed in Chapter 2, and perhaps other phenomena.

## 6.2 The Basic Story

To illustrate how a BiGLA-based evolutionary story might help explain a language’s transition from Stage 1 to Stage 3 and beyond, let us restrict the focus of the present discussion to the distribution of anaphoric expressions and leave the discussion of Principle C-type effects and the distribution of R-expressions aside for the moment. Let us further suppose that we are dealing with a language like Old English, wherein the relevant inventory was limited to pronouns and *pro+self* forms, and restrict our attention to only two types of inputs, core conjoint transitive clauses and core disjoint ones.

We can start with a set of frequencies that might correspond to a dialect I’ll call ‘Keenan’s Old English’, wherein (just as in the survey of OE sources circa 750-1154, per Keenan (2001, 15)) 18% of the locally conjoint object pronouns are *self*-marked, the rest bare. For simplicity, I’ll assume that the ratio of core-disjoint versus core-conjoint transitive clauses is 49:1. And for good measure – though this is also not from Keenan’s data and is merely done to illustrate a point – I will assume that 18% of locally disjoint objects are also *self*-marked, for reasons of contrast, emphasis, or something else.<sup>1</sup>

---

<sup>1</sup>As noted, Keenan argues that the OE *self* morpheme lost its contrastive meaning in A-positions by virtue of Decay and then gained a reflexive meaning by virtue of Antisynonymy. Levinson advocates a slightly different picture wherein the contrastive properties of *self* are not seen as having disappeared, but are analyzed as pragmatic byproducts in the first place which became more narrow in the sense that *pro+self* came to indicate only referential contrast (to stereotype) rather than any other kind.

By leaving out any mention of values for a feature like, say,  $\pm$ *contrastive* from the various inputs, I am essentially simulating a post-Decay/pre-Antisynonymy scenario, in the Keenan senses, and avoiding the questions about various other kinds of contrast. This is done only for the purpose choosing a simple starting point, and the picture could obviously be enriched to describe competition between reflexive interpretations and, say,

(6.1) *Frequencies per ‘Keenan’s OE’*

	<i>pro</i>	<i>pro+self</i>	<i>%marked</i>
<i>conjoint</i>	1.64%	.36%	18%
<i>disjoint</i>	80.36%	17.64%	18%

A child-learner exposed to the frequencies of Keenan’s OE would thus be learning a grammar in which there was no correlation between whether a form was marked and whether the input associated with that form was conjoint or disjoint. Rather, the *self*-marking in (6.1) is non-differential.<sup>2</sup>

As in Chapter 5, I will simulate the spirit of Zeevat’s *Bias<sub>int</sub>* in the best way I can think of by assuming that a learner’s grammar consists partly of a ‘bias-calculator’ of sorts, i.e., a comprehensive pool of codistributional constraints that refer to specific pairs of form-meaning types and which – per the assumptions of Wilson (2001) and Buchwald et al. (2002) – are relevant to both the generative evaluation procedure and the interpretational one, and are thus subject to adjustment in both hearer- and speaker-modes of learning.

(6.2) *Bias constraints*<sup>3</sup>

*\*self,co*: *Self*-marked pronouns are not locally conjoint.

*\*self,dis*: *Self*-marked pronouns are not locally disjoint.

*\*pro,co*: Bare pronouns are not locally conjoint.

*\*pro,dis*: Bare pronouns are not locally disjoint.

In addition to *\*Struct*, a nascent learner will need to learn ranking values for these constraints based on the training corpus frequencies he is exposed

---

emphatic interpretations. But for reasons of scope and lack of access to any good corpus data in those regards, I must leave such complications as a matter of further research.

<sup>2</sup>Whether such a scenario really ever obtained at any point in the diachronic history of English is unclear, though intuitively very doubtful, it seems. Faltz (1985, 328) notes that “potentially reflexive uses are one typical context for use of an emphatic”, and, as noted, Levinson’s line of explanation about the evolution of reflexives from emphatics would lead us to believe that *self*-marking or emphatic marking of any other kind in a Stage 1 language would never be truly non-differential with respect to locally conjoint and locally disjoint clauses. However, I assume for illustration that this is the case, if only to level the relevant playing field and make the experiment a bit more convincing.

<sup>3</sup>Constraints very similar to the ones in (6.2) were first suggested to me by Henk Zeevat.

to and the BiGLA. For the sake of illustration, let us ignore *\*Struct* for the moment and consider only the learning effects that the various form-meaning pairs would have on the various bias constraints.

Recall that, in speaker-mode, a learner observes a pair  $\langle f, m \rangle$  and, using the observed meaning  $m$  as an input, generates a hypothetical output. Where that hypothesis is wrong, the learner demotes all and only those constraints which favor the hypothesis and promotes all and only those which favor the observed output. Where we take ‘ $\uparrow$ ’ to mean ‘gets promoted’ and ‘ $\downarrow$ ’ to mean ‘gets demoted’, we can outline the speaker-mode learning effects of the various types of learning data as in the table below.

(6.3) *Speaker-mode learning effects (bias constraints)*

Observed pair	hyp.	* <i>pro, dis</i>	* <i>self, dis</i>	* <i>pro, co</i>	* <i>self, co</i>
$\langle pro, dis \rangle$	pro+self	$\downarrow$	$\uparrow$		
$\langle self, dis \rangle$	pro	$\uparrow$	$\downarrow$		
$\langle pro, co \rangle$	pro+self			$\downarrow$	$\uparrow$
$\langle self, co \rangle$	pro			$\uparrow$	$\downarrow$

On the other hand, in hearer-mode, a learner observes a pair  $\langle f, m \rangle$  and, using the observed expression  $f$  as an input, generates a hypothetical output. Again, where that hypothesis is wrong, the learner demotes all and only those constraints which favor the hypothesis and promotes all and only those which favor the observed meaning. The hearer-mode learning effects of the various types of pairs in our corpus can be summarized as follows.

(6.4) *Hearer-mode learning effects (bias constraints)*

Observed pair	hyp.	* <i>pro, dis</i>	* <i>self, dis</i>	* <i>pro, co</i>	* <i>self, co</i>
$\langle pro, dis \rangle$	co	$\downarrow$		$\uparrow$	
$\langle self, dis \rangle$	co		$\downarrow$		$\uparrow$
$\langle pro, co \rangle$	dis	$\uparrow$		$\downarrow$	
$\langle self, co \rangle$	dis		$\uparrow$		$\downarrow$

There are a couple of things we can note.

First of all, because  $*self,co$  and  $*self,dis$  together represent a general dispreference for *self*-marked forms and  $*pro,co$  and  $*pro,dis$  combine to represent a general reluctance toward bare pronouns (and since all outputs are either bare pronouns or *self*-marked ones, but never both) it will hold as a matter of logic that any grammar we consider will always remain such that, ranking value-wise, the sum of  $*self,co$  and  $*self,dis$  is the opposite of the sum of  $*pro,co$  and  $*pro,dis$ .

(6.5) *General fact 1*

- a.  $*self,co + *self,dis = x$
- b.  $*pro,co + *pro,dis = -x$

In a similar way, since the constraints  $*self,co$  and  $*pro,co$  together generally disfavor conjoint interpretations and  $*self,dis$  and  $*pro,dis$  collectively represent an aversion to disjoint ones (and since all interpretations are either disjoint or conjoint, but never both) it will also hold generally that the sum of the values of  $*self,co$  and  $*pro,co$  is the opposite of the sum of  $*pro,dis$  and  $*self,dis$ .

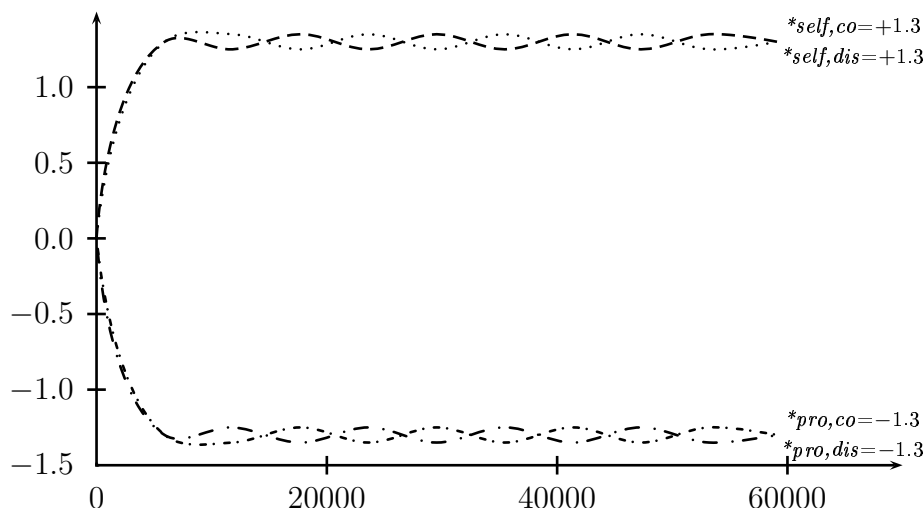
(6.6) *General fact 2*

- a.  $*self,co + *pro,co = y$
- b.  $*self,dis + *pro,dis = -y$

As for what we can expect to see from a BiGLA-learned grammar whose learning data were the frequencies in (6.1):

With respect to speaker-mode learning, we know that since the percentage of disjoint inputs that get *self*-marked outputs and the percentage of conjoint inputs that get them is exactly the same (viz. 18%), we can expect that, insofar as speaker-mode learning is concerned, there will be little or no difference between the ranking values of the constraints  $*self,co$  and  $*self,dis$ , and likewise for  $*pro,co$  and  $*pro,dis$ . On the other hand, because of the asymmetry between bare pronouns and *self*-marked forms in general – the former greatly outnumbering the latter – we know that speaker-mode learning will rank  $*self,co$  and  $*self,dis$  significantly higher than  $*pro,co$  and  $*pro,dis$ .

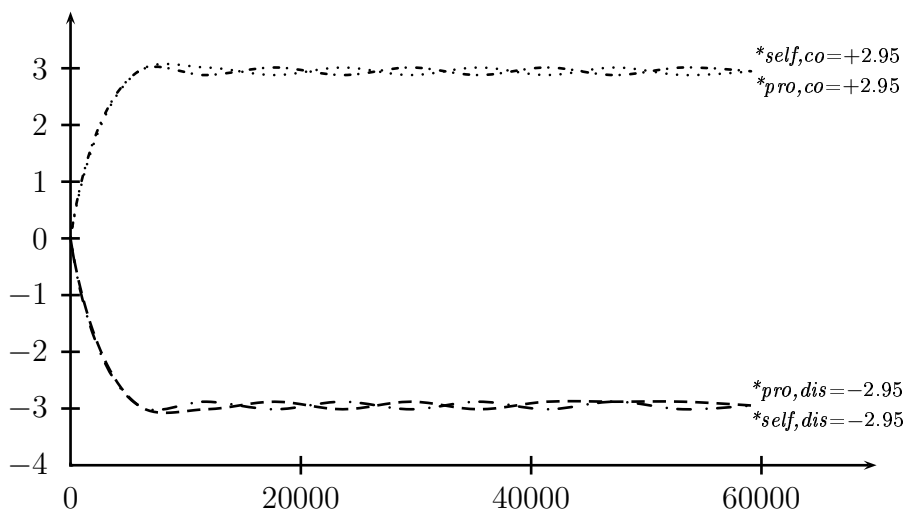
(6.7) *Speaker-mode learning curves (bias constraints)*



In hearer-mode the situation will be very different. Because 98% of the pairs in the corpus are disjoint and only 2% are conjoint, hearer-mode learning will result in the net promotion of *\*self,co* and *\*pro,co* and the net demotion of *\*self,dis* and *\*pro,dis*. On the other hand, because conjoint and disjoint inputs get *self*-marked outputs an equal percentage of the time in the training corpus, hearer-mode learning will essentially not recognize a distinction in rank between *\*self,co* and *\*pro,co*, nor between *\*self,dis* and *\*pro,dis*.

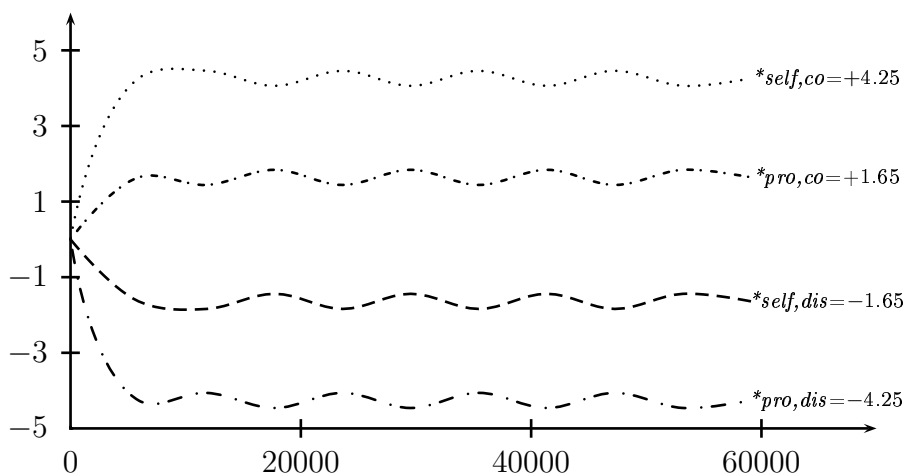
In this way, hearer-mode learning will be ‘pulling apart’ exactly the two pairs of constraints that speaker-mode learning is ‘holding together’ (viz. *\*self,co*/*\*self,dis* and *\*pro,co*/*\*pro,dis*). On the other hand, hearer-mode learning is also holding together the two pairs of constraints that speaker-mode learning is pulling apart (viz. *\*self,co*/*\*pro,co* and *\*self,dis*/*\*pro,dis*).

(6.8) *Hearer-mode learning curves (bias constraints)*



The effects of bidirectional learning will thus be a compromise of sorts between the two learning modes. As such, after sixty-thousand observations of Keenan's OE, given a grammar consisting only of the bias constraints in (6.2), the learning curves would ideally look like the ones below.

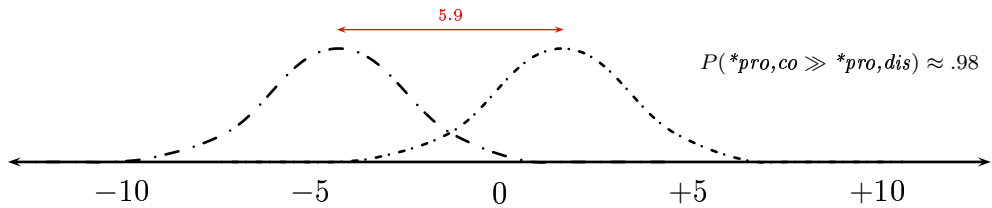
(6.9) *Bidirectional learning curves (bias constraints)*



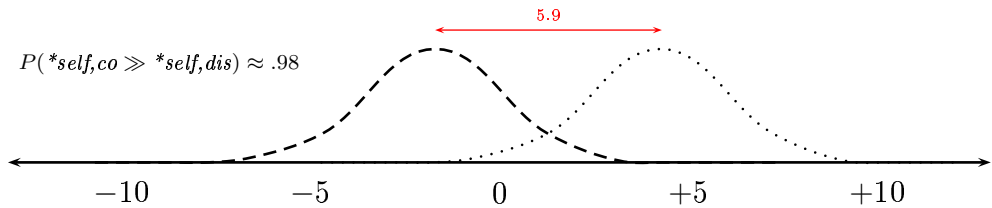
Even at a glance, one might be able to see that this compromise is a successful one in that the grammar in (6.9) reflects the frequencies in (6.1) accurately.

Firstly, with respect to the interpretational preferences, the grammar reflects the expected, general preference for disjoint interpretations.

(6.10) *Interpretational optimization for bare pronouns, per (6.9)*



(6.11) *Interpretational optimization for self-marked forms, per (6.9)*

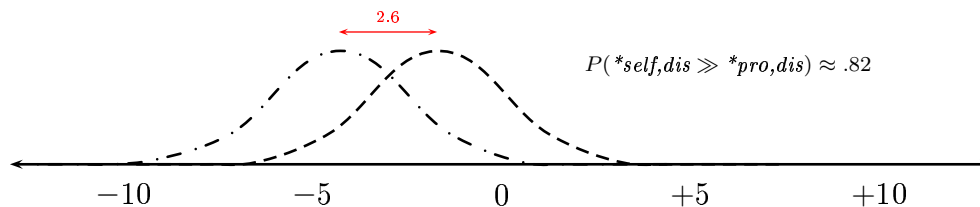


Note that what is shown in (6.10) and (6.11) is basically a stochastic version of the DRP. But rather than stipulating it as a pragmatic presumption à la Farmer & Harnish (1987), a pragmatic implicature toward stereotypicality à la Levinson (1991, 2000), or a ‘derivative of world-knowledge’ à la Huang (1994, 2000), a statistically sensitive bidirectional learning algorithm like the BiGLA can provide a functional explanation for how and why DRP-like effects came to be. The preference for disjoint interpretations is derived directly from an asymmetry in the training corpus and the application of hearer-mode learning to bias constraints which ‘record’ that asymmetry.

The generative preferences exhibited in the training corpus are reflected accurately in the learned grammar as well. Firstly, by ranking and *\*pro,dis* above *\*self,dis* the grammar reflects a preference for bare pronouns, given locally disjoint inputs.

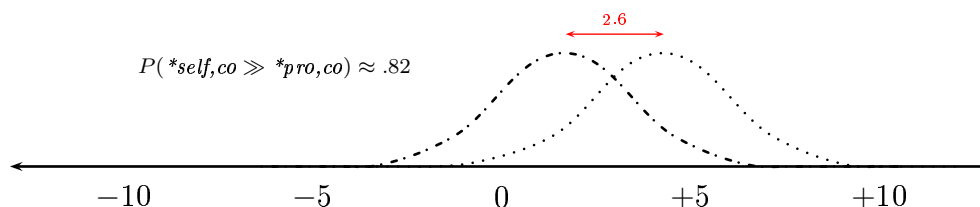


(6.12) *Generative optimization for disjoint inputs, per (6.9)*



And the same holds true with respect to conjoint inputs, since  $*self,co$  dominates  $*pro,co$ .

(6.13) *Generative optimization for conjoint inputs, per (6.9)*



Note that the distance between the constraints in (6.12) is equal to the distance between the constraints in (6.13) and that the distance between the constraints in (6.10) is equal to the distance between the constraints in (6.11). It is in this way that the grammar is able to reflect the fact that, in the original corpus, disjoint inputs were just as (un)likely to get *self*-marked outputs as conjoint inputs were and that *self*-marked forms were just as (un)likely to be associated with conjoint inputs as bare pronouns were. In other words: all the constraints of the grammar can be and have been learned in such a way that the interpretational frequencies and output frequencies of the training corpus are represented accurately; there is no irresolvable conflict or dissonance between the two modes of learning.

For exactly this reason, the grammar in (6.9) is ‘stable’; a learner who has internalized such a grammar and who was exposed to learning data like that in (6.1) would learn virtually nothing, since he would not draw an incorrect hypothesis very often in either of the two learning modes. For this reason, we can expect that the frequencies manifested in this learner’s own speech will be a very close, if not totally exact, replica of the training corpus frequencies in (6.1).

Adding a markedness constraint like  $*Struct$  to represent some universal force of structural economy causes this picture to change significantly.

To understand why, first consider the various learning effects pertaining to *\*Struct*.

(6.14) *Learning effects on \*Struct*

Observed Pair	hyp.	<i>*Struct</i>
$\langle pro, dis \rangle$	self	↑
$\langle self, dis \rangle$	pro	↓
$\langle pro, co \rangle$	self	↑
$\langle self, co \rangle$	pro	↓

Note that because it militates against *self*-marked outputs, *\*Struct* will get promoted iff *\*self,co* or *\*self,dis* is promoted in speaker-mode and will get demoted in speaker-mode iff *\*pro,co* or *\*pro,dis* is promoted in speaker-mode. For this reason, we know that *\*Struct* will always remain such that its ranking value is exactly the sum of the ranking values of *\*self,co* and *\*self,dis* (and thus will be the opposite of the sum of the values of *\*pro,co* and *\*pro,dis*).<sup>4</sup>

(6.15) *General fact 3*

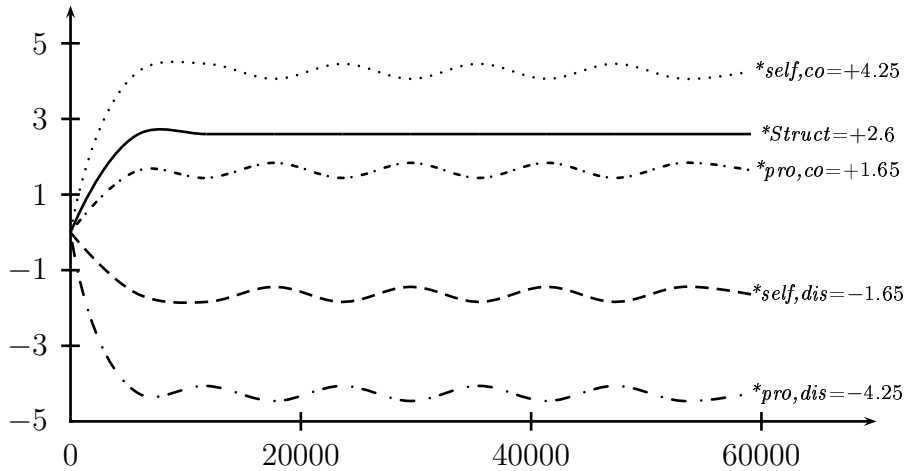
$$*Struct = *self,co + *self,dis$$

To see why this will be relevant to determining how a grammar with both bias constraints and markedness constraints will converge, just consider what would happen if we were to calculate the value of *\*Struct* based on General fact 3 and the ranking values in (6.9). We know that the ranking value for *\*Struct* would be the sum of the values of *\*self,co* and *\*self,dis*, which is, in this case, +2.6. We would have:

---

<sup>4</sup>Though *\*Struct* is not promoted or demoted in hearer-mode whereas the bias constraints are, this does not interfere with the fact stated here, since whenever, say, *\*self,dis* is promoted in hearer-mode, *\*self,co* will get demoted. For this reason, the net promotion of *\*Struct* in hearer-mode is indeed still always equal to the net promotion of *\*self,co* + *\*self,dis*, i.e., zero.

(6.16) Grammar in (6.9), with *\*Struct*



This grammar would *not* reflect the output frequencies of the original learning corpus accurately. The general reason: generative optimization in a grammar with both bias constraints and markedness constraints will be determined not only by the ranking values of bias constraints, but also by how the markedness constraints are ranked among them.

In the simplest hypothetical example: where two constraints  $C_1$  and  $C_2$  say ' $\alpha$ ' and one constraint  $C_3$  says ' $\neg\alpha$ ', then to get 50-50 optionality between  $\alpha$  and  $\neg\alpha$  we cannot simply rank all three constraints equally. (This would give us 67%-33% in favor of  $\alpha$ .) Rather,  $C_1$  and  $C_2$  need to be about one unit lower than  $C_3$  to get free optionality. This is because in order for  $\neg\alpha$  to be the optimal output for a particular evaluation,  $C_3$  must outrank *both*  $C_1$  and  $C_2$ . Jäger & Rosenbach (2003) call this effect *ganging-up cumulativity* – each constraint is relevant to the evaluation regardless of its ranking value.<sup>5</sup>

With respect to the case at hand, (ignoring bidirectional optimization for the moment) the probability that a *self*-marked output is the optimal output for, say, a conjoint input is now no longer equal to the probability that *\*pro,co*

<sup>5</sup>More specifically, for any set of ranked constraints  $C_1 \gg \dots \gg C_n$ , where  $r_i$  is the ranking value of  $C_i$  and  $N$  is the standard normal distribution:

$$P(C_1 \gg \dots \gg C_n) = \int_{-\infty}^{+\infty} dx_1 N(x_1 - r_1) \int_{-\infty}^{x_1} dx_2 N(x_2 - r_2) \int_{-\infty}^{x_{n-1}} dx_n N(x_n - r_n)$$

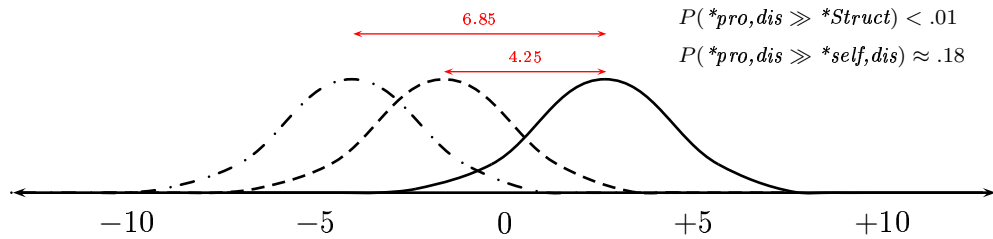
Cf. Jäger (2003b) and Jäger & Rosenbach (2003) for more details.

outranks *\*self,co*, but rather to the probability that *\*pro,co* outranks both *\*self,co* and *\*Struct*. Thus, a grammar like the one under consideration this needs to converge in a way such that the markedness constraint and the bias constraints ‘share the labor’ in the prevention of *self*-marked forms.

This is significant, since, I will argue, it is exactly this ‘ganging-up cumulativity’ effect of markedness constraints and certain bias constraints that will ultimately be responsible for the fact that a marked-form-for-marked-meaning pattern is the only evolutionary stable target for Keenan’s Old English (or anything like it) in the context of bidirectional, GLA-style learning.

To use (6.16) as an example, consider what the generative optimization in that grammar would look like.

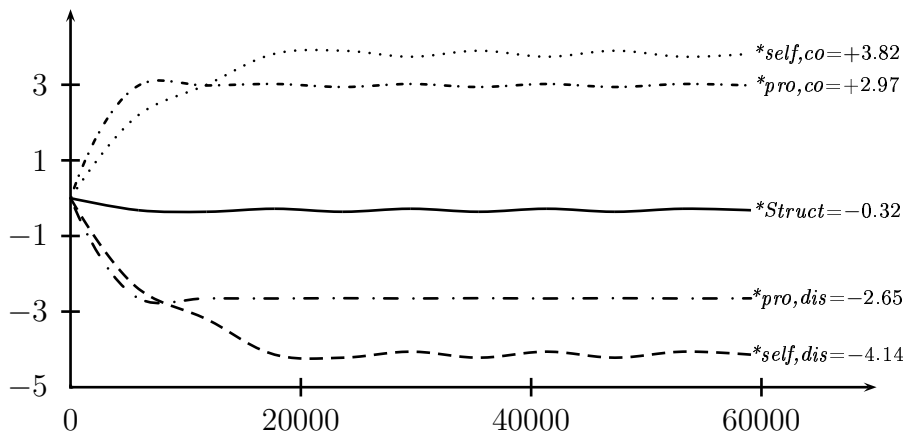
(6.17) *Generative optimization for disjoint inputs, per (6.16)*



A learner whose hypothetical grammar was the one in (6.16) would almost never hypothesize *self*-marked forms as optimal outputs for locally disjoint inputs. Likewise, a speaker with an internalized grammar like the one in (6.16) would almost never use *self*-marked forms, especially not for disjoint objects. Such a grammar would obviously be unstable in light of the training corpus, since 18% of all NPs were marked there.

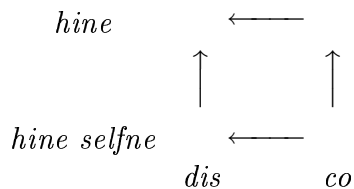
For this reason, it would be hugely unlikely for a BiGLA-learner observing a corpus with frequencies as in (6.1) to ever find himself in a state where his hypothetical grammar looked like the one in (6.16). Instead, feeding BiGLA with sixty-thousand observations drawn at random based the frequencies of Keenan’s OE produced the learning curves below.

(6.18) *Bidirectional learning curves (first generation)*



As with previous cases, this grammar exhibits a general preference for unmarked forms and a general preference for disjoint interpretations and is thus again a case in which the pair  $\langle pro, dis \rangle$  is a strongly bidirectionally optimal pair and  $\langle pro+self, co \rangle$  is a weakly bidirectionally optimal one.

(6.19)

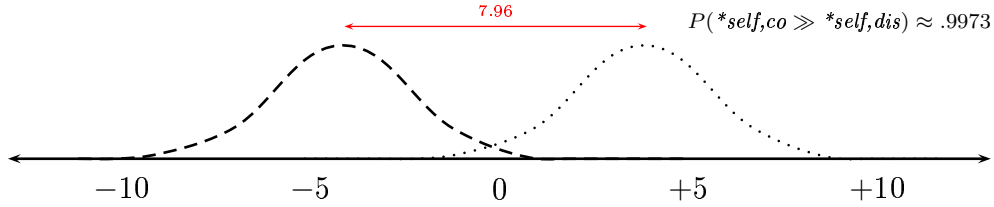


Importantly though, this grammar differs greatly from the one in (6.9) in that the grammar in (6.18) is not stable.

There are two main areas of instability.

The first unstable aspect of the grammar in (6.18) is the interpretational optimization with respect to marked forms. Consider the constraints relevant to that evaluation:

(6.20) *Interpretational optimization for self-marked forms, per (6.18)*



Given such rankings, *self*-marked forms are optimally interpreted as conjoint less than 1% of the time. The grammar under consideration is thus not stable in hearer-mode, since in the training corpus 2% of *self*-marked forms are conjoint.

The grammar is unstable in speaker-mode as well. This can be seen just by looking at the actual output frequencies for (6.18). They were:

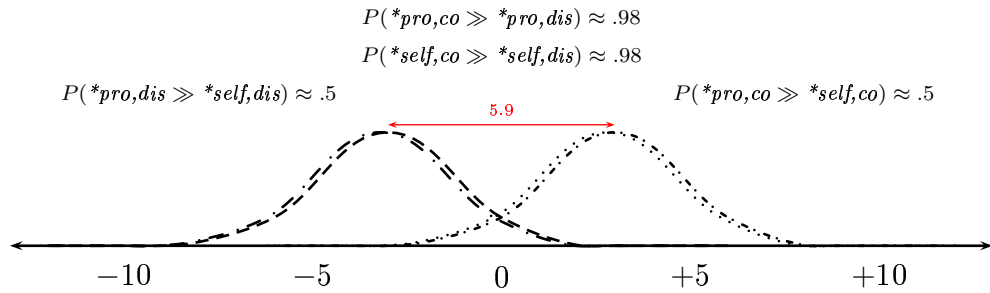
(6.21) *Frequencies (first generation)*

	<i>pro</i>	<i>pro+self</i>	%marked
<i>co</i>	1.26%	0.74%	37%
<i>dis</i>	77.64%	20.36%	20.8%

The across-the-board 18% *self*-marking has not held up at all in the learned grammar. Rather, a cross-generational change has occurred such that conjoint inputs are now more likely as compared to disjoint ones to be expressed as *self*-marked forms. In other words, marked forms have gravitated toward marked meanings.

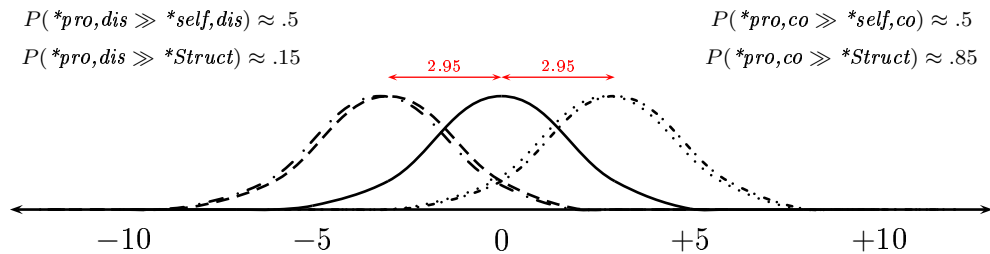
To appreciate why a BiGLA-learned grammar based on the bias constraints in (6.2) and the constraint *\*Struct* suffers from such instability when faced with the frequencies in (6.1), consider once again a scenario wherein only hearer-mode learning was taking place and *\*Struct* was not present. We would have:

(6.22) *Hearer-mode learning (bias constraints)*



As noted above, when we consider bias constraints alone, hearer-mode learning does not register any generative preferences. Thus in a hypothetical grammar like (6.22), bare pronouns and *self*-marked ones would occur in free variation. However, if the markedness constraint *\*Struct* is considered, hearer-mode learning will have very significant, if accidental, generative effects. There will, of course, be no adjustment of *\*Struct* in hearer-mode, but the mere fact that it is there will have serious consequences. For example, with *\*Struct* added to the picture above and ranked appropriately at zero, we would get:

(6.23) *Hearer-mode learning (bias constraints + markedness constraints)*



The grammar in (6.23) would predict that *self*-marked forms are the optimal candidate for conjoint inputs around 40-50% of the time. On the other hand, *self*-marked forms would be the optimal candidate for disjoint inputs only about 12-15% of the time.<sup>6</sup> The marked form, though still dispreferred generally, is being repelled from the statistically prevalent meaning with much greater force as compared to the statistically rare one due to the

<sup>6</sup>These are just my very rough estimates. I leave the actual calculation to the reader, cf. fn. 5, though precision in this regard is not at all crucial for the argument.

effect of hearer-mode learning and the effects of ‘ganging-up cumulativity’. In this way, the marked-forms-for-marked-meanings pattern can be seen as a consequence of four things: (a) bias constraints (b) markedness constraints (c) the mechanics of the GLA and (d) the bidirectional application of those mechanics.

Of course, the actual learned grammar in our experiment is not one in which 40-50% of conjoint objects and only 12-15% of disjoint ones got *self*-marked. The difference between the actual learned grammar in (6.18) and the hypothetical one in (6.23) is just the effect of speaker-mode learning.

Given the frequencies in (6.1), the task of speaker-mode learning is to find a set of ranking values whereby the probability that *\*pro,co* outranks both *\*self,co* and *\*Struct* is .18 and whereby the probability that *\*pro,dis* outranks both *\*self,dis* and *\*Struct* is also .18. This could be done in any number of ways, but none of them would be able to preserve the accuracy of the interpretational evaluation procedure. It is simply impossible for BiGLA to learn the five constraints under consideration in a way that perfectly reflects both the interpretational frequencies and the output frequencies of the training corpus. In other words, given the frequencies of the training-corpus and the constraints under consideration, stability in one learning mode entails instability in the other mode. Hearer-mode and speaker-mode are in conflict, and some compromise had to be reached.

If one were to describe the ‘strategy’ according to which that compromise was reached, one could say the reasoning went roughly as follows:

*Speaker-mode* would ‘reason’ that if the constraints governing the generative optimization for disjoint inputs (viz. *\*pro,dis* and *\*self,dis* and *\*Struct*) converge so that output frequencies in the training corpus *for disjoint inputs* are perfectly reflected, then the learner’s speech will be 98% accurate (minimally). Thus, speaker-mode learning will strongly tend to converge in this way.

*Hearer mode* will ‘reason’ that if the constraints governing interpretational optimization for bare pronouns (viz. *\*pro,dis* and *\*pro,co*) converge so that the interpretational frequencies in the training corpus *for pronouns* are perfectly reflected in the learner’s interpretations, then the learner’s interpretations will be 82% accurate (minimally). Thus, hearer-mode learning will strongly tend to converge in this way.

Of course, there is really no reasoning or strategizing going on in the learning here. Rather, something very much like this ‘strategy’ is just logically the most probable outcome of bidirectional GLA-learning.



Note that one constraint is never mentioned in the convergence strategy: *\*self,co*. Speaking loosely, given the frequencies in (6.1), speaker-mode learning would take a hearer-mode-only grammar like (6.23) and push *\*self,co* up while pushing *\*self,dis* down, in hopes of compensating for the ‘accidental’ generative effects hearer-mode learning had (due to the presence of *\*Struct*). However, it will do so in a way that more accurately reflects the generative optimization for *disjoint* inputs rather than the generative optimization for *conjoint* inputs (i.e., it gives *\*self,dis* roughly the ‘right’ ranking and *\*self,co* the ‘wrong’ one – it’s ranked too low to make the right generative predictions about the teacher’s speech), per the speaker-mode convergence strategy above. The accuracy of the generative optimization for *conjoint* inputs is being ‘sacrificed’ because they are *rarer*. (Of course, again, there is no real ‘sacrifice’ going on here that involves conscious decision-making, just an imbalance of learning effects and consequences of that.)

Thus, it is exactly because *conjoint* inputs were the rarer type of input that the new asymmetry in percentages of *self*-marked forms will be heavy on the *conjoint* side, not light on the *disjoint* side.

In this way, it is exactly the application of bidirectional learning to a set of bias constraints and the way that certain bias constraints ‘gang-up’ with markedness constraints that is responsible for initiating the marked-form-for-marked-meaning pattern that we are seeing here.

This general pattern and the factors behind it can, I think, provide part of the explanation we are looking for with respect to the transition of a Stage 1 language into a Stage 3 language.

### 6.3 The Next Generation

Of course, the grammar in (6.18) would still only qualify as a Stage 1 grammar. There are no reflexives here, only a pronoun, which is preferably interpreted as *conjoint*, and a *self*-marked pronoun, which is even more likely to be interpreted as *disjoint*. (One might dare call it an ‘emphatic’.)

However, we can expect the new asymmetry that has shown up in the first-generation learner’s corpus frequencies to have important consequences for future generations. Per the ILM, the student who produces a greater percentage of *self*-marked outputs for *conjoint* inputs than he does for *disjoint* inputs will eventually become a teacher to the next generation and thus a second-generation learner will be exposed to a training corpus in which the

tendency to *self*-mark locally conjoint pronouns is greater than the tendency to mark locally disjoint ones.

Insofar as speaker mode-learning in generation-two is concerned, there will not be any significant cross-generational changes. The first-generation learner's grammar reflects his own speech accurately in terms of the number of *self*-markers he employs and where he puts them; the effects of bidirectional optimization on his speech were negligible.<sup>7</sup>

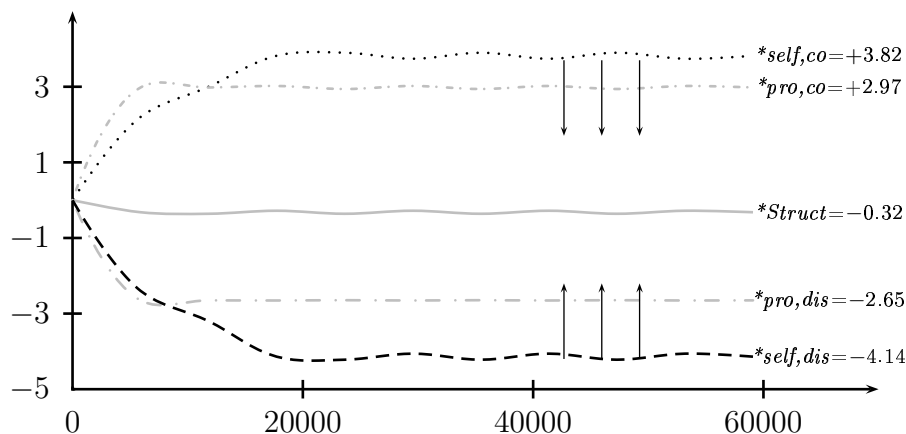
On the other hand, hearer-mode learning effects will be very significant. The constraints *\*self,co* and *\*self,dis* are, as we saw, very unstable in the first-generation learned grammar. The ranking of *\*self,co* and *\*self,dis* in the first-generation learner's grammar not only makes inaccurate interpretational predictions for his teacher's speech, but it makes even worse predictions for his own speech; the nearly-eight unit difference between those two constraints in (6.18) does not even come close to corresponding to what is now a .74%/22%  $\approx$  3.4% chance that a *self*-marked form is conjoint (per the first-generation learner's output (6.21)); eight units is way too much.

For this reason, second-generation hearer-mode learning – especially the observation of *self*-marked conjoint objects – will effect the most cross-generational changes. We can expect that hearer-mode learning in generation-two will always be trying to converge in a way that 'improves upon' his teacher's grammar in terms of interpretational preferences by ranking the constraints in a way that more accurately reflects his teacher's interpretational frequencies (i.e., the interpretational frequencies of (6.21)). In this case, this could only be accomplished by lowering the ranking value *\*self,co* and raising the ranking value of *\*self,dis* (compared to the previous generation). Moreover, since there are twice as many *self*-marked conjoint objects in the first-generation learner's speech as compared to the original Keenan's OE training corpus, the second generation will have twice the amount of learning data favoring this direction of change as his teacher did.

---

<sup>7</sup>The only noticeable effect is that of the about 2% of disjoint objects are *self*-marked due to bidirectional optimization. That effect will be short-lived and insignificant, though.

(6.24) *Cross-generational hearer-mode adjustments (generations 1-2)*

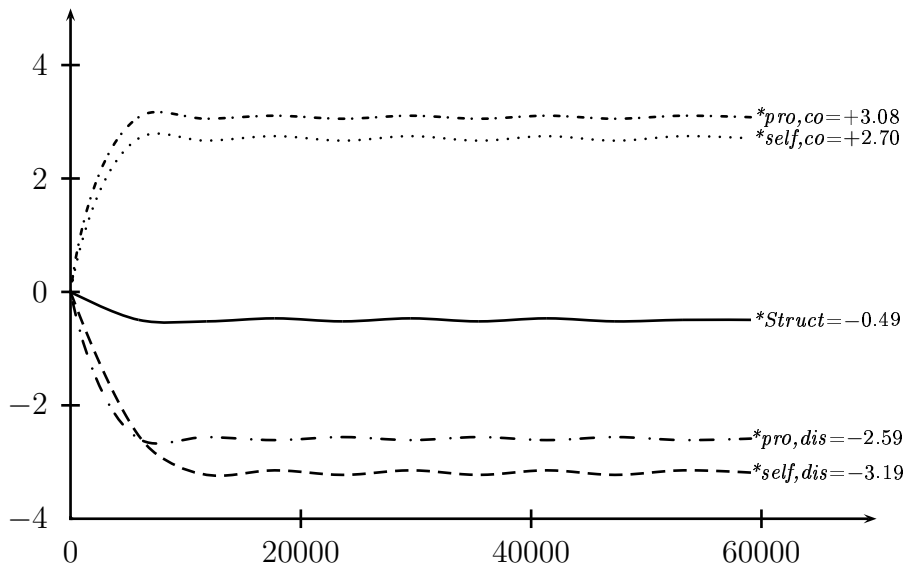


However, because we are assuming that the same set of bias constraints operates in both the generative and interpretational evaluation procedures, each ‘corrective’ measure in hearer-mode will have consequences for the generative evaluation procedure as well. In particular, by lowering the ranking value of *\*self,co* so that the probability of interpreting *self*-marked forms as conjoint is higher, the learner also commits to a ranking that demands expressing conjoint inputs as *self*-marked outputs more often. So, by ‘improving’ his teacher’s grammar with respect to its reflection of the teacher’s own interpretational frequencies, a learner necessarily learns a constraint ranking that does not perfectly predict the teacher’s speech.

In this way, we can predict for what Zeevat & Jäger (2002) called a “self-reinforcement” of the marking pattern. Our grammar has begun to ‘chase its tail’: each adjustment of the constraints made by the hearer-mode learning of one generation will entail generative consequences that require the next generation to make even further hearer-mode adjustments.

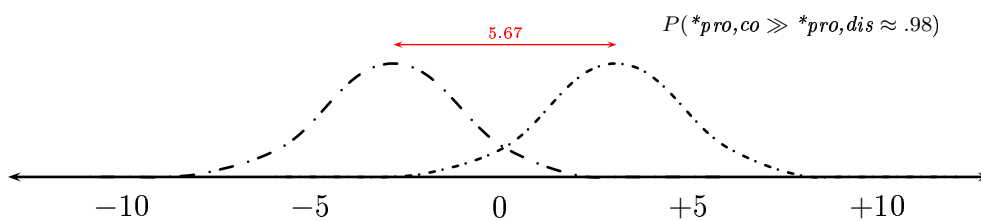
We can feed BiGLA with sixty-thousand inputs drawn at random based on the frequencies in (6.21) to see what another turn in the cycle actually looks like.

(6.25) *Bidirectional learning curves (second generation)*

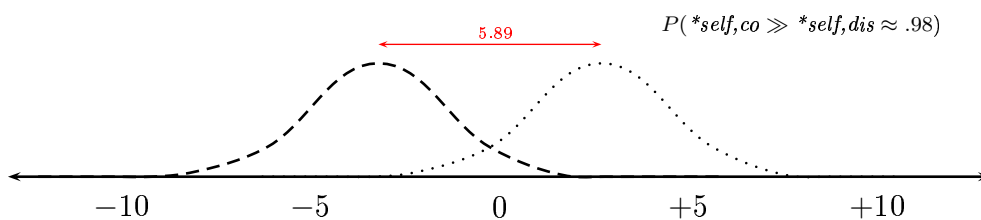


Consider the constraints relevant to interpretational optimization in the grammar above.

(6.26) *Interpretational optimization for bare pronouns, per (6.25)*



(6.27) *Interpretational optimization for self-marked forms, per (6.25)*



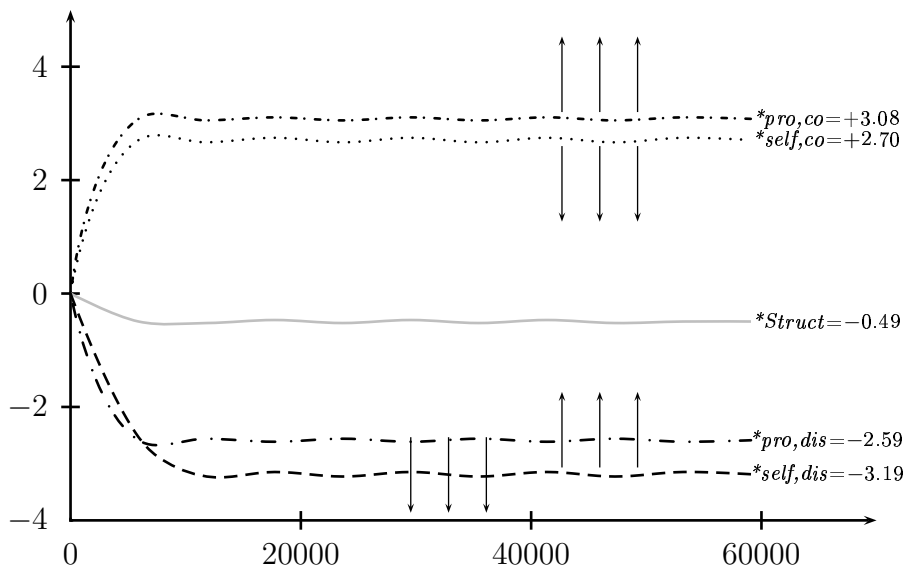
About 2% of *self*-marked forms and 2% of bare pronouns will now be interpreted as conjoint according to the rankings in (6.25). Note that this grammar would reflect fairly accurately the interpretational optimization for the *original* training corpus frequencies in (6.1). Of course, the frequencies in (6.1) are *not* the training corpus on which the grammar in (6.25) was based. But, like most evolved characteristics, the properties of this grammar are a bit behind the times. Moreover, the hearer-mode adjustments that have taken place will, as noted above, result in an increase in  $\langle pro+self, co \rangle$  pairs. Even more hearer-mode adjustment will thus be needed in the third generation, since the percentage of *self*-marked forms that are conjoint in speech of the second generation is not 2%, as in the original Keenan's OE training corpus, nor  $.74\%/22\% \approx 3.4\%$  as in the first generation's speech, but now  $1.12\%/20.26\% \approx 5.5\%$ , per (6.28), below.

(6.28) *Frequencies (second generation)*

	<i>pro</i>	<i>pro+self</i>	%marked
<i>co</i>	.88%	1.12%	56%
<i>dis</i>	78.94%	19.14%	19.5%

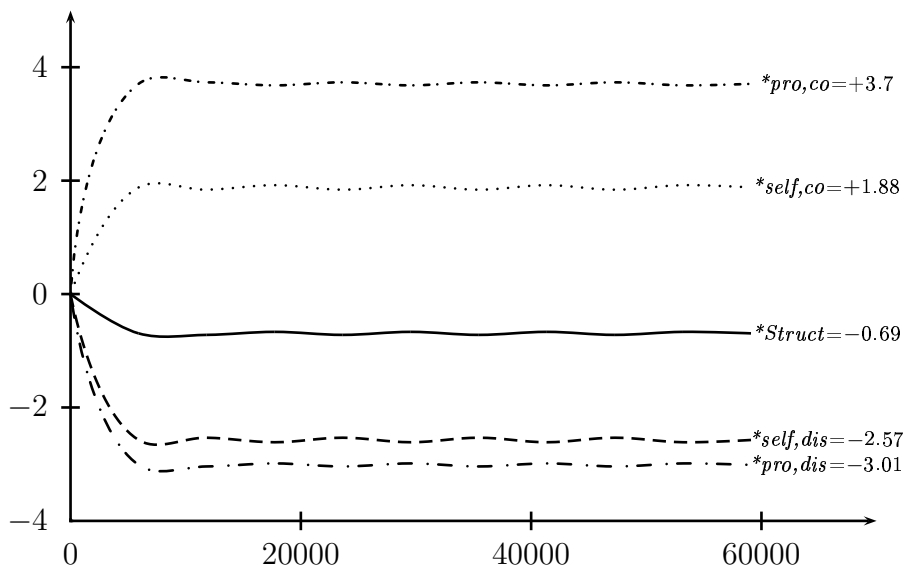
The odds that a *self*-marked form is conjoint have more than doubled in just two generations of iterated learning. And thus the 'tail-chasing' continues. For a third-generation learner, it will again be hearer-mode learning that is responsible for the cross-generational adjustments. Specifically, the greater percentage of *self*-marked conjoint objects will cause  $*self,co$  and  $*self,dis$  to be learned as closer together than they were in his teacher's grammar. Moreover, for exactly the same reason,  $*pro,co$  will get cross-generationally promoted and  $*pro,dis$  will get cross-generationally demoted, since there are ever increasing odds that a bare pronoun is *not* conjoint.

(6.29) *Cross-generational hearer-mode adjustments (generations 2-3)*



The actual learning curves of a BiGLA-learned grammar based on the output frequencies of the second generation looked as follows.

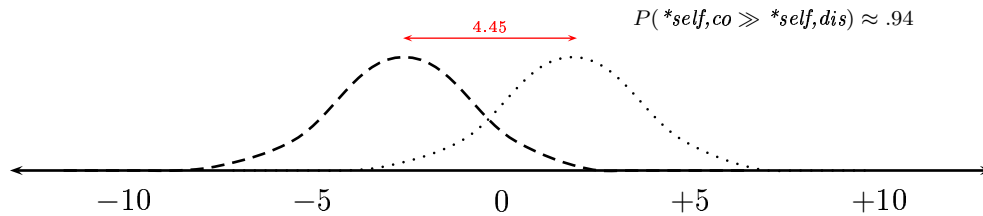
(6.30) *Bidirectional learning curves (third generation)*



Note that though *self*-marked pronouns are still (in the most probable case) optimally interpretable as disjoint, the likelihood that *self*-marking will solicit a conjoint interpretation is now significantly greater than the likelihood that a pronoun will. (We might compare this to Levinson’s ‘Stage 2’.)

For a speaker who has internalized the grammar in (6.30), the effects of bidirectional optimization on his speech will be significant. The reason: *\*self,dis* and *\*self,co* are now close enough together that there is now about a 6% chance that *\*self,dis* will outrank *\*self,co* and thus 6% that a *self*-marked form will be optimally interpreted as conjoint.

(6.31) *Interpretational optimization for self-marked forms, per (6.30)*



Thus, in about 6% of the cases where a third-generation speaker wants to express a conjoint input, he will use a *self*-marked form, regardless of whether the unidirectional generative optimization favors this or not.<sup>8</sup> Thus, to use Jäger’s terminology, we will start to see ‘pragmatic’ *self*-marking on top of ‘structural’ *self*-marking.

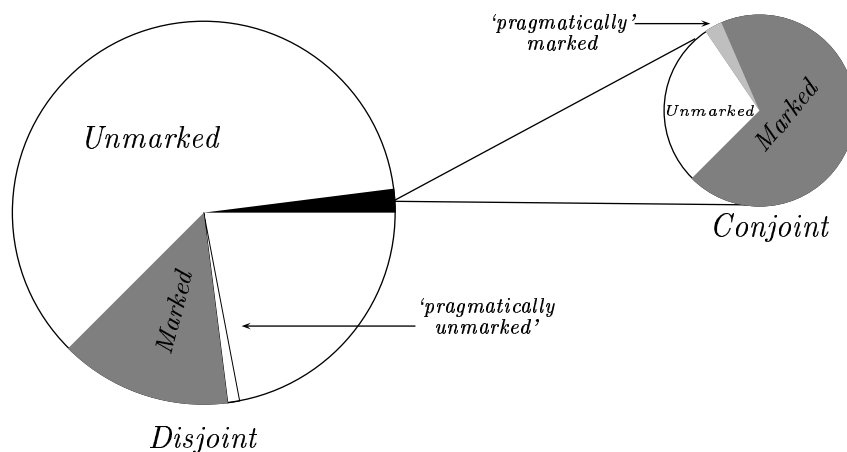
(6.32) *Frequencies (third generation)*

	<i>pro</i>	<i>pro+self</i>	%marked
<i>co</i>	.56%	1.44%	72%
<i>dis</i>	83.88%	14.12%	14.5%

We can estimate that about 3% of conjoint inputs here are *self*-marked due to bidirectional optimization and not the unidirectional generative optimization itself. Moreover, we can estimate that about 1% of disjoint inputs are *not* marked for exactly the same reason. That is, where unidirectional generative optimization favors a *self*-marked form for a disjoint input, that optimization is occasionally overridden by bidirectional optimization and a bare pronoun must be used; we might call this ‘pragmatic *un*-marking’.

<sup>8</sup>Provided of course that that *\*pro,dis* does not also outrank *\*pro,co*, but this would be extremely improbable.

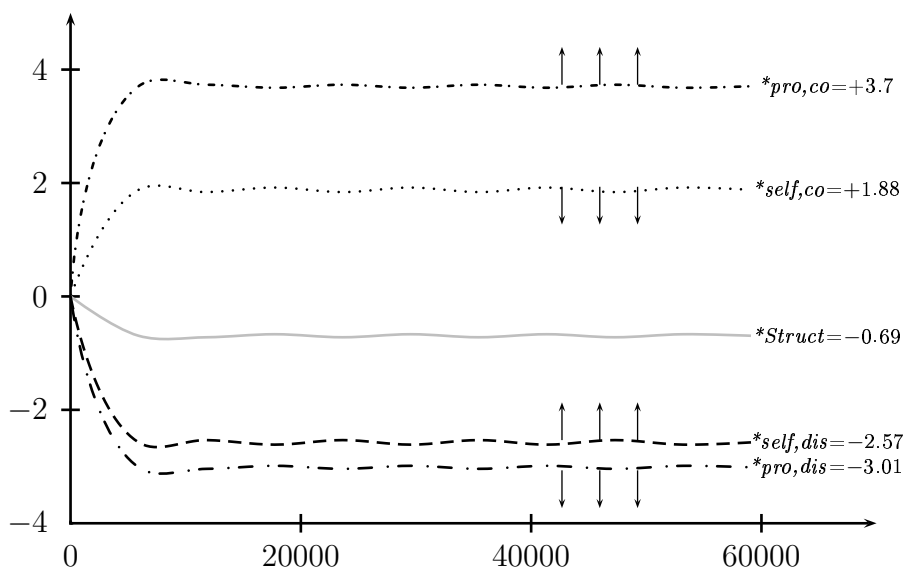
(6.33) *Marked versus unmarked (third generation)*



The effects of bidirectional evaluation will have significant effects on the future of the grammar in similar ways that we have already seen in the accounts of differential case marking, per (Zeevat & Jäger, 2002), (Cable, 2002), et al., discussed in Chapter 5. Generally speaking, the effects of bidirectional optimization in generation  $n - 1$  are learned by generation  $n$  as if they were the effects of simple unidirectional, generative optimization. In particular, though about 4% of the  $\langle pro+self, co \rangle$  pairs in (6.32) are there because of bidirectional optimization, a learner learning a grammar based on those frequencies will treat them as if they were the effects of unidirectional optimization. Most importantly, the observation of the pragmatically marked pairs will induce hearer-mode learning of the next generation to learn the rankings of  $*self, co$  and  $*self, dis$  as being closer together than they were in his teacher's grammar. Similarly, the observation of the pragmatically unmarked pairs will induce the learner to learn the rankings of  $*pro, dis$  and  $*self, dis$  as being further apart (per speaker-mode learning) and  $*pro, dis$  and  $*pro, co$  as being further apart as well (per hearer-mode), compared to the previous generation.



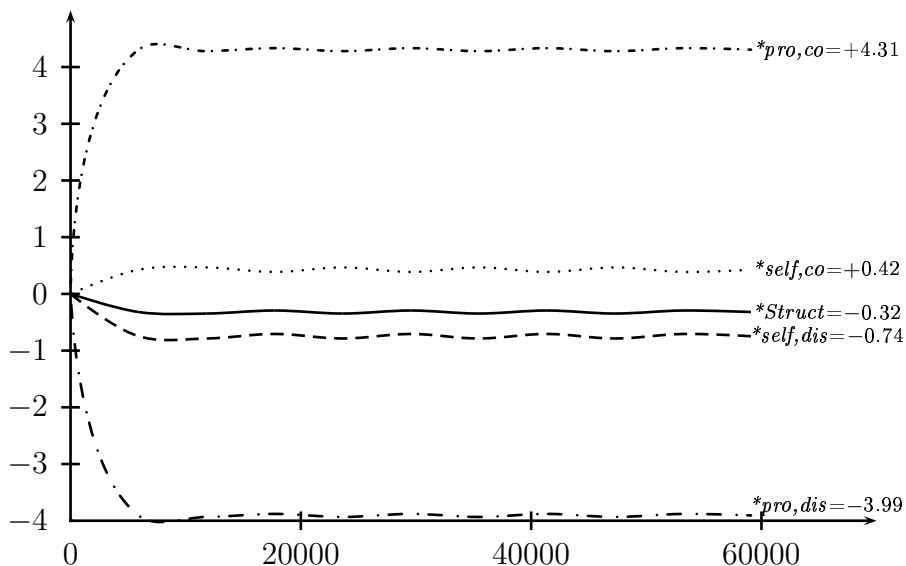
(6.34) *Cross-generational consequences of bidirectional optimization (generations 2-3)*



Note that in this way the marking-strategy begins to reinforce itself in a second way, since – just like in the Zeevat & Jäger (2002) picture – the more significant bidirectional optimization is, the more significant it will become. In particular, per the adjustments depicted in (6.34), *self*-marked forms will become more likely optimally interpreted as conjoint. For that reason, the generative effects of bidirectional optimization will become even greater, since bidirectional optimization will overrule unidirectional generative optimization more often, creating more ‘pragmatically marked’ conjoint objects and more ‘pragmatically unmarked’ disjoint ones.

We can fast-forward ten more generations of iterated learning to observe the net effect of the two types of self-reinforcement. A thirteenth-generation descendant of a Keenan’s OE speaking ancestor could learn a grammar like the one in (6.35).

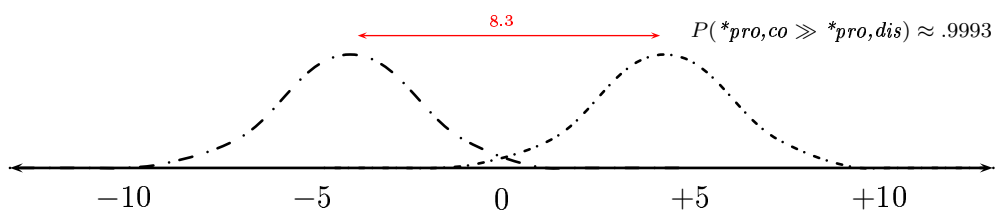
(6.35) *Bidirectional learning curves (thirteenth generation)*



Note that something like Chomsky’s ‘Principle B’ is almost fully instated in this grammar.

In the first place, the odds that a bare pronoun will be optimally interpreted as locally conjoint is more than a thousand-to-one.

(6.36) *Interpretational optimization for bare pronouns, per (6.35)*



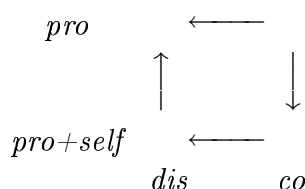
And secondly, because *\*pro,co* has been learned as about four units higher than both *\*self,co* and *\*Struct*, generative optimization will heavily favor *self*-marked forms for conjoint inputs. The actual output frequencies of were:

(6.37) *Frequencies (thirteenth generation)*

	<i>pro</i>	<i>pro+self</i>	%marked
<i>co</i>	.14%	1.86%	93%
<i>dis</i>	95.54%	2.46%	2.5%

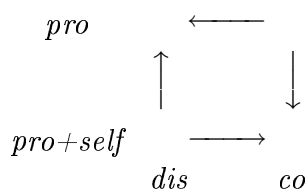
Note that while the grammar under consideration still possesses only one strongly bidirectionally optimal pair, in the Blutner (2000b) sense, it is a grammar where the weakly optimal pair exhibits unidirectional optimality on one side (in the most probable case).

(6.38)



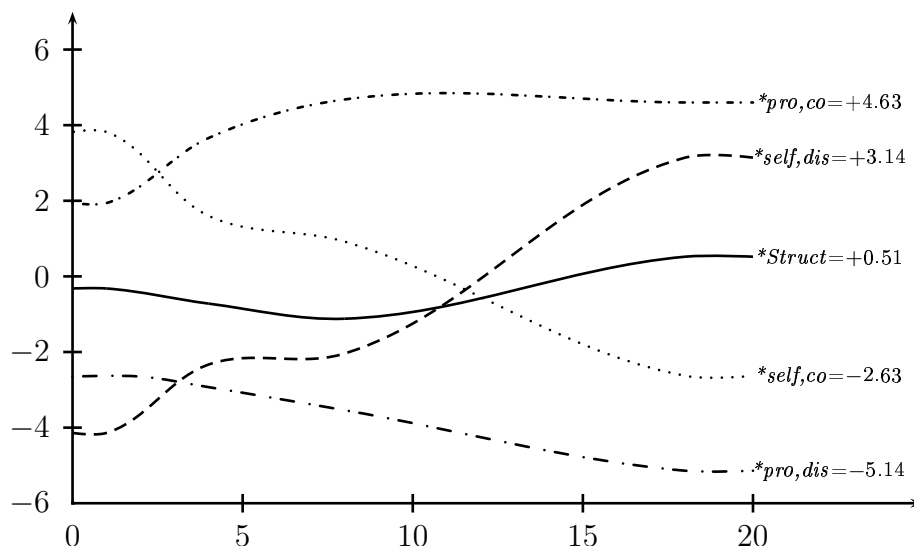
Moreover, in just one more generation of iterated learning, the constraint *\*self,dis* overtook *\*self,co*, and Keenan's OE became a language wherein there was strong bidirectional optimality relations between both pairs, since *pro+self* will now be optimally interpretable as conjoint (probably). We might see this as a passage from Stage 2 to Stage 3, though generation-fourteen would still be an 'early Stage 3' grammar, since there is still plenty more grammaticalization to be done.

(6.39)



All told, the evolutionary transition of 'Keenan's Old English' in Modern Standard English, per twenty generations of iterated BiGLA-learning looked roughly as below.

(6.40) *Evolution of 'Keenan's Old English' (generations 1-20)*



Note the S-shaped curves reminiscent of the trajectory of historical change discussed by Kroch (1989). The hearer-mode/speaker-mode 'dissonance' responsible for these curves is further reminiscent of the view of Kroch (1989), Pintzuk (1991), Santorini (1992), and Kroch & Taylor (1997) that speakers during language change should be described by multiple grammars in competition with one another. On the picture above, it is clear that the competing forces are hearer-mode and speaker mode learning. Both forces are now in harmony with one another, since all constraints in the grammar can now be learned in such a way the interpretational frequencies and the output frequencies of the relevant training corpus can be reflected accurately.

The evolved grammar strictly follows the Principle A and B patterns of Modern Standard English.

(6.41) *Frequencies (twentieth generation)*

	<i>pro</i>	<i>pro+self</i>	% <i>marked</i>
<i>co</i>	0%	2%	100%
<i>dis</i>	98%	0%	0%

Note that the general marked-forms-for-marked-meanings pattern predicted in the picture above has very little to do with the initial frequencies

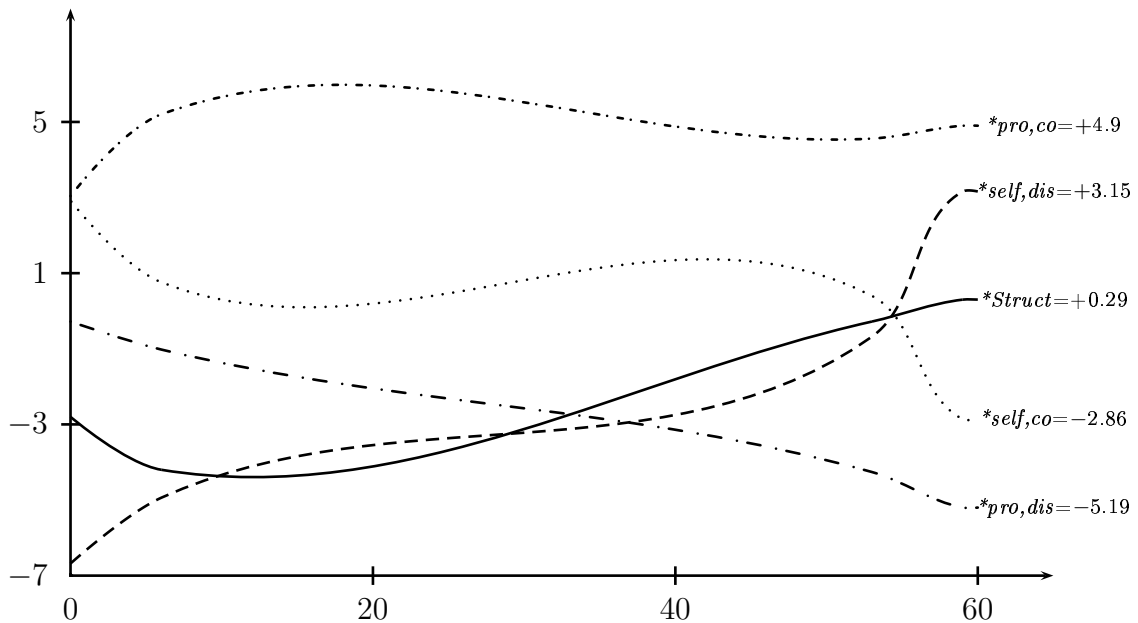
in the original training corpus. Case in point, we can assume, as before, that 18% of conjoint objects are *self*-marked but, unlike before, 82% of disjoint objects were (for whatever reason) *self*-marked as well, a semi-backwards version of our Keenan’s Old English.

(6.42) *Frequencies ‘Semi-backwards Keenan’s Old English’*

	<i>pro</i>	<i>pro+self</i>	%marked
<i>co</i>	1.64%	.36%	18%
<i>dis</i>	17.64%	80.36%	82%

The results after 60 generations of iterated bidirectional learning:

(6.43) *Evolution of ‘Semi-backwards Keenan’s Old English’ (60 generations)*

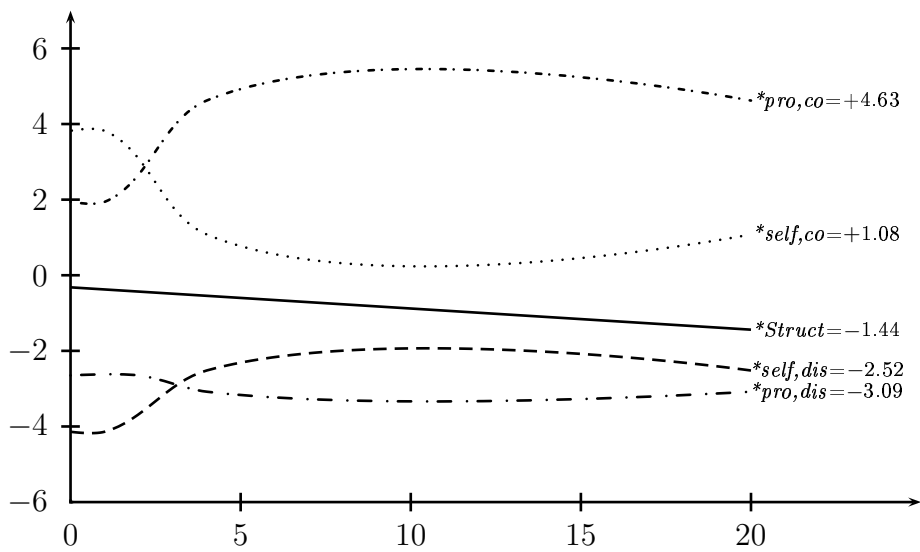


The end-result is practically identical to the previous case. In short: because there are still many more disjoint pronouns than conjoint ones, hearer-mode learning will push *\*pro,co* high. This will have the ‘accidental’ generative consequence that marked (conjoint) meanings will tend toward being expressed with *self*-marked forms, not pronouns. These new  $\langle pro+self,co \rangle$  pairs will have the obvious effect on the hearer-mode learning of subsequent

generations. Then, eventually, bidirectional optimization will gradually prohibit expressing unmarked meanings with the marked forms.

Note that it is indeed the complicit influence of bidirectional learning *and* bidirectional optimization that is responsible for the change shown above. To illustrate this point, consider the results of evolutionary iterated bidirectional learning of Keenan’s OE wherein bidirectional optimization did *not* play a role.

(6.44) *Evolution of ‘Keenan’s OE’ sans bidirectional optimization*



The output frequencies for such a grammar were:

(6.45) *Frequencies, per (6.44)*

	<i>pro</i>	<i>pro+self</i>	<i>%marked</i>
<i>co</i>	.28%	1.72%	86%
<i>dis</i>	80.08%	17.9%	18.3%

Marked forms have generally gravitated toward marked meanings as a result of bidirectional learning. However there has been little change with respect to the frequencies for ‘unmarked’, i.e., disjoint, meanings compared to the original training corpus.

In this case, then, it is fair to say that *bidirectional learning* and the interaction of bias constraints with markedness constraints is primarily responsible for getting us the marked-form-for-marked-meaning pattern. On the other hand, it is *bidirectional optimization* that typically ensures the marked-form-*only-for*-marked-meaning pattern.

I will suggest below, however, that, in certain cases, a marked-form-(only)-for-marked-meaning pattern can arise where bidirectional optimization is only partially relevant or not immediately relevant to generative optimization at all. Examples might include cases of what Keenan called ‘Pattern Generalization’ and/or cases of differential SE/SELF distribution, as in Dutch, or other cases of multiple reflexivizing strategies. The extension of the account above to such cases depends partly on the idea that bias constraints can also ‘gang-up’ with another type of constraint, namely *faithfulness constraints*, and that this can also play a role in the stability of a bidirectionally-learned grammar.

## 6.4 Pattern Generalization

Above I have tried to show how, given a comprehensive set of ‘bias constraints’ and a universal structural constraint, a learning model like Jäger’s (2003a) Bidirectional Gradual Learning Algorithm can predict for the marked-forms-for-marked-meanings strategy noted by Shannon (1948), Atlas & Levinson (1981), Horn (1984), and Blutner (2000b), et al. I argued that that general pattern could be seen as a result of bidirectional learning and the ‘ganging-up’ interaction between bias constraints and markedness constraints, and that this can corroborate to some degree, yet clarify considerably, the claims of Levinson (2000), et al. to the effect that certain binding phenomena are the manifestation of a marked-forms-for-marked-meanings pragmatic strategy.

There are a wide range of issues that I have left untouched so far. E.g.: I have neglected any mention of languages like Greek, Malagasy, or Hungarian, discussed in Chapter 2, which appear to be more obedient to the Thematic Hierarchy Condition than to Principles A and B. The issue of ‘long-distance Anaphors’ that has troubled standard binding theories, discussed in Chapters 2 and 3 is unaccounted for as well. And of course, Principle C and surrounding issues discussed in Chapter 2 have not been discussed since then. I will return briefly to these issues in the next section, though some of the actual

experimental work related to them must remain an area of further research.

One issue, though, that is acutely interesting in the context of the discussion above and in the previous chapter is the issue of what Keenan (2001) referred to as ‘Pattern Generalization’. Keenan invoked Pattern Generalization for the purpose of filling in what is a blind spot for his ‘Antisynonymy’ (and also for Levinson’s M-principle, Blutner’s weak bidirectional optimality, and the like). Specifically, Pattern Generalization is meant to account for the spread of *self*-marking to local person objects. Such marking conveys no change in meaning, since *me* and *myself* always refer to the same individual.

I think that at least some types of Pattern Generalization might be explained – or, at the very least, described more precisely – in a framework of bidirectional evolutionary learning, per Jäger (2003a). What is more, I think that a bidirectional-learning-based account of Pattern Generalization might be extendable to cases of involving multiple, discriminatory reflexive marking strategies like that of Dutch, which was a focal point of Reinhart & Reuland’s (1993, et al.) semantic alternative to Chomsky’s Binding Theory (1986, et al.), for just as *me* and *myself* are synonymous, it seems that *zich* and *zichzelf* are as well.

In the experiment conducted above, I made no distinction between local and non-local person pronouns, and the implicit assumption was that all arguments were third person, since it was being taken for granted that pronouns and *pro+self* forms were all potentially ambiguous, unlike *me* and *myself*.

However, we can set up a more fine-grained experiment wherein the relevant distinction is drawn and, I believe, show how the effects of iterated bidirectional learning and bidirectional optimization can nevertheless still predict for the evolutionary development of a categorical marking strategy like *self*-marking in Modern Standard English, wherein all locally conjoint objects are marked regardless of their person orientation.

In order to represent the force pertaining to the fact that first person pronouns are not interpreted as third person (and vice versa), we can assume the existence of a general faithfulness constraint, *Faith*, that demands faithfulness with respect to person features, semantic content, etc.

To illustrate how an evolutionary story based on bidirectional learning might be able to account for Pattern Generalization, let us imagine that we have a language much like our original ‘Keenan’s Old English’ such that 18% of all pronouns are *self*-marked whether the object is conjoint or disjoint, and whether it is of local or non-local person orientation. Assume also, as



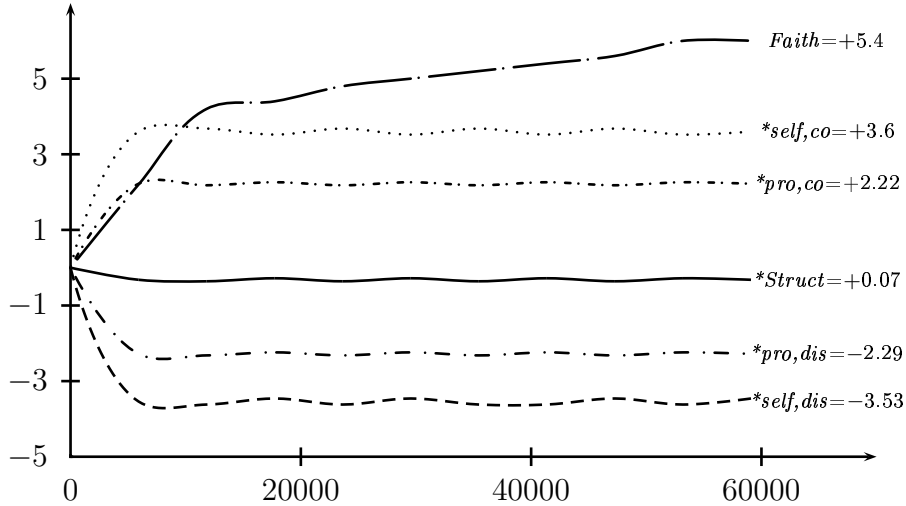
before, that the ratio of disjoint to conjoint transitive clauses is 49:1. And, for the purpose of incorporating the local/non-local person distinction into the experiment, assume that half of the arguments of transitive clauses were local person (let's say first-person) and the other half were third person. We could restrict our attention to the inputs and outputs shown in (6.46):

(6.46) *Revised frequencies for 'Keenan's OE'*

	I hit me	I hit myself	He hit him	He hit himself	He hit me	He hit myself	I hit him	I hit himself
1 hit 1	.82%	.18%	0	0	0	0	0	0
3 <sub>i</sub> hit 3 <sub>j</sub>	0	0	20.09%	4.41%	0	0	0	0
3 <sub>i</sub> hit 3 <sub>i</sub>	0	0	.82%	.18%	0	0	0	0
3 hit 1	0	0	0	0	20.09%	4.41%	0	0
1 hit 3	0	0	0	0	0	0	40.18%	8.82%

If we assume the exact same regimen of constraints as before, plus *Faith*, then bidirectional learning will just have the usual effect: Hearer-mode will learn to generally disfavor conjoint interpretations. This will mean that, as before, two pairs of bias constraints – *\*self,co* and *\*pro,co* on the one hand and *\*self,dis* and *\*pro,dis* on the other – are spread apart from one another to reflect ‘DRP-like’ interpretational preferences. In this way, hearer-mode learning will have the same ‘accidental’ generative effects by virtue of the ‘gang-up’ interaction of the markedness constraint, *\*Struct*, and the bias constraints *\*self,dis*, and *\*self,co*, just as in the previous experiment. Namely, the grammar will develop a tendency to express marked meanings with marked forms.

(6.47) *Bidirectional learning curves (first generation)*



We can use (6.46) as an ancestor corpus and execute 60 generations of iterated BiGLA learning to show that the lack of referential ambiguity between, say, *me* and *myself* does not greatly hinder the pattern that showed up in the earlier experiment. The output frequencies of an evolved grammar after sixty generations of iterated learning bear testament to this.

(6.48) *Frequencies of Revised Keenan's OE (after sixty generations)*

	I hit me	I hit myself	He hit him	He hit himself	He hit me	He hit myself	I hit him	I hit himself
1 hit 1	0	1%	0	0	0	0	0	0
3 <sub>i</sub> hit 3 <sub>j</sub>	0	0	24.5%	0	0	0	0	0
3 <sub>i</sub> hit 3 <sub>i</sub>	0	0	0	1%	0	0	0	0
3 hit 1	0	0	0	0	24.5%	0	0	0
1 hit 3	0	0	0	0	0	0	49%	0

Pattern Generalization was complete in this experiment after about twenty-five or thirty generations. The marking of first person conjoint object pronouns often lagged slightly behind the third person ones by a few generations, but this is to be expected, since bidirectional optimization will potentially manifest blocking effects in cases where both subject and object are third person, but not in other cases. (And, if we believe Keenan (2000), it might

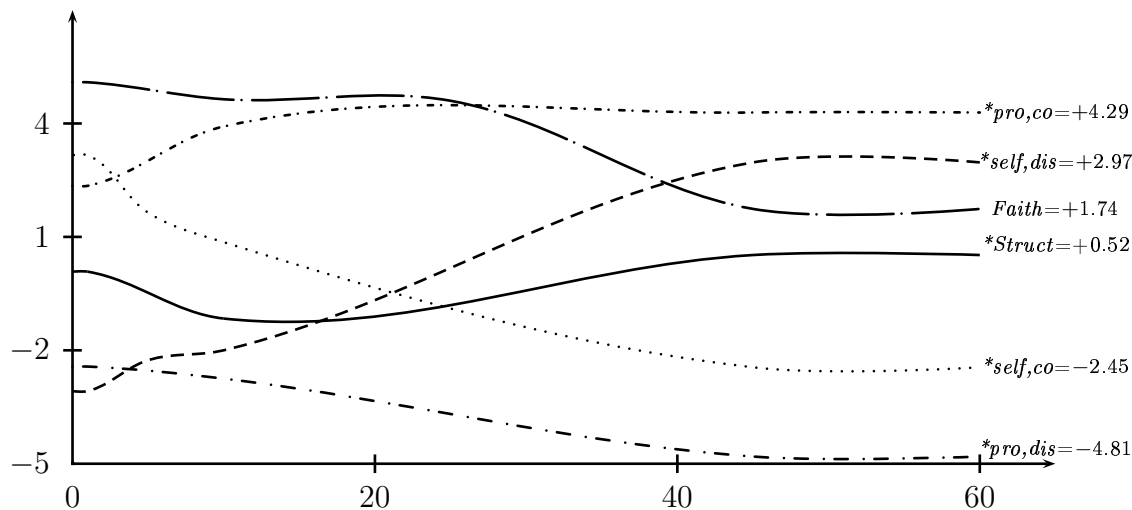
actually be quite like the historical picture that we see when we look at the spread of *self*-marking in Middle English.)

In terms of the evolution of this grammar, what happened is that, once again, bidirectional learning has resulted in the cross-generational demotion of *\*self,co* and *\*pro,dis* and the cross-generational promotion of *\*self,dis* and *\*pro,co*. Bidirectional optimization, though irrelevant for most of the cases, it still relevant to some of the cases. And, when it is relevant, it will have the same effects it had in the previous experiment, namely, the cross-generational demotion of *\*self,co* and the cross-generational promotion of *\*self,dis*, as well as the cross-generational demotion of *\*pro,dis* and the cross-generational promotion of *\*pro,co*. Both the learning trends related to bidirectional learning and those due to bidirectional optimization will be, as in the previous experiment, ‘self-reinforcing’, and the effects will be magnified with each generation. Since these ‘learning trends’ are just the demotion and promotion of particular bias constraints, and since the four bias constraints we are considering are indeed relevant arguments of all person orientation, their cross-generational promotion or demotion will have the usual consequences (i.e., forcing *self*-marking for conjoint objects, pronouns for disjoint ones), regardless of whether bidirectional optimization shows its hand or not. The results are thus virtually no different from those in the earlier experiment.<sup>9</sup>

---

<sup>9</sup>It is worth noting that we could enrich the pool of bias constraints so that the grammar recognized person orientation and its relation to *self*-marking and was able to form biases on this dimension as well. In my own experience, experiments which included constraints such as, say, *\*self,3* and the like, in addition to the ones already present, did not produce the kind of non-fuzzy end-results shown above, though they did tend to exhibit a strong tendency toward the marked-forms-for-marked-meanings pattern, just as a result of hearer mode learning effects. I must leave as an area of further research questions of how to sensibly integrate various sets of bias constraints involving multiple dimensions of input characteristics into an experiment and how to represent that some dimensions seem much more relevant than others.

(6.49) *Evolution of Revised Keenan's OE (generations 1-60)*



Note the clever positions of *\*pro,co* and *\*self,dis*, which are now ranked even higher than *Faith*. This is about as much grammaticalization as one could hope for! Moreover, I think it illustrates something else important about the types of grammars that tend to be stable in iterated bidirectional learning experiments and it is a point that, I will cautiously claim below, might help account for another area of data, namely the differential SE/SELF distribution in Dutch and perhaps other differential marking strategies which are not clear-cut cases of ‘pragmatic-to-structural’ marking, in the Jäger (2003a) senses, i.e., they are not cases in which bidirectional optimization ‘evolved into’ unidirectional optimization, since bidirectional optimization does not play an immediate role in the generative optimization of any generation.

The significant cross-generational demotion of *Faith* in the evolution of the grammar in (6.49) does not represent that *Faith* has been demoted at any point in time. Presumably, *Faith* has *never* been demoted, since presumably no learner ever observes a pair that violates it. Rather, *Faith* is simply being *promoted less* in later generations. And the reason for this is exactly that, in the later generations, learners are drawing less unfaithful hypotheses during the learning process. At least part of the reason for this is that as the pattern of *self*-marking generalizes to all conjoint objects, a learner becomes more likely to be faithful (in his hypotheses). In other words, Pattern Generalization ‘fosters’ faithfulness. To see how, consider the following small thought

experiment.

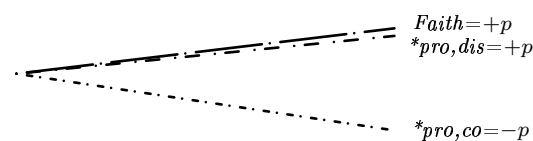
Suppose that a newborn learner was born into a speech community wherein only non-local person objects were obligatorily *self*-marked but Pattern Generalization was incomplete. With his constraints ranked at zero, suppose the learner observes an expression *I hit me*, with the obvious meaning. Suppose that, per chance, his speaker-mode hypothesis is correct, but his hearer-mode hypothesis is incorrect. Above we have considered five inputs, four of which would qualify as incorrect. The various learning effects of those four are as below.

(6.50) *Hearer-mode learning effects per 'I hit me'*

Observed pair	hypothesis	<i>Faith</i>	<i>*pdis</i>	<i>*pro</i>	<i>*self,dis</i>	<i>*self,co</i>
<i>I hit me</i>	'3 <sub>i</sub> hit 3 <sub>j</sub> '	↑	↑	↔		
'1 hit 1'	'3 <sub>i</sub> hit 3 <sub>i</sub> '	↑			*	*
	'3 hit 1'	↑	↑	↓		
	'1 hit 3'	↑	↑	↓		

If we assume that the hypothesis was one of the three that does result in some learning effect on the bias constraints then we know that *Faith* and *\*pro,dis* go up while *\*pro,co* goes down (by  $p$ , where  $p$  is the plasticity value).

(6.51) *(Partial) rankings after first observation (case 1)*



Now suppose the learner is confronted with a second observation: *I like him*, with the obvious meaning.

It is not unlikely that the learner would draw an incorrect hearer-mode hypothesis at this point. In particular, the odds are pretty good that he will hypothesize that *I like him* gets a conjoint interpretation such as '1 hit 1' or '3<sub>i</sub> hit 3<sub>i</sub>'. The reason: *\*pro,dis* is equally ranked with *Faith*. Of course,

*\*pro,co* is still relevant, so the faithful hypothesis is still more likely, but the highest ranked bias constraint and *Faith* are in conflict here.

On the other hand, imagine another newborn learner born into a speech community wherein Pattern Generalization was almost or totally complete.

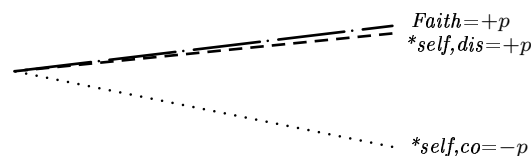
Suppose his constraints were ranked at zero and he observed his first input/output pair: an expression *I hit myself*, with the usual meaning. Suppose now that, in speaker-mode, his hypothesis was correct but in hearer mode his hypothesis was incorrect. The possibilities are:

(6.52) *Hearer-mode learning effects per 'I hit myself'*

Observed pair	hypothesis	<i>Faith</i>	<i>*pro,dis</i>	<i>*pro,co</i>	<i>*self,dis</i>	<i>*self,co</i>
<i>I hit myself</i> <i>hit(1,1)</i>	'3 <sub>i</sub> hit 3 <sub>j</sub> '	↑			↑	↓
	'3 <sub>i</sub> hit 3 <sub>i</sub> '	↑				
	'3 hit 1'	↑			↑	↓
	'1 hit 3'	↑			↑	↓

Again, assume that the hypothesis was one of the three that would effect some learning effect with respect to the bias constraints. We know that *Faith* and *\*self,dis* get promoted and *\*self,co* gets demoted.

(6.53) *(Partial) rankings after first observation (case 2)*



Now suppose the learner encounters a second observation: *I hit him*.

This learner is less likely to interpret this expression unfaithfully than the first imaginary learner was. The reason: the present learner has not ranked *\*pro,dis* equally with *Faith* like the first learner did. There is thus less conflict between the bias constraints and the faithfulness constraint. Moreover, if our second learner were to encounter a second observation of *He hit himself*, he is not only more likely to draw a faithful hypothesis because of his earlier promotion of *Faith*, but also because of the earlier promotion of *\*self,dis*.

Thus, whereas the first learner's bias constraints were in conflict with *Faith*, the second learner's bias constraints are working more in concert with *Faith*.

In this way, a marking pattern that has generalized so that it is categorically differential on some dimension – e.g., conjoint/disjoint – will induce learning effects of bias constraints that can reinforce faithfulness.

This is basically the reason *Faith* takes a sharp turn down in the evolutionary experiment shown in (6.49). Where differential *self*-marking (including Pattern Generalization) is nowhere near categorical, *Faith* will be violated often early on in the learning process (of each generation) and must therefore get ranked very high. On the other hand, where a marking strategy is a clean 'split-system' and where Pattern Generalization has thus become categorical, bias constraints will work in tandem with *Faith*, allowing it to remain ranked low.

Note that this is yet another case of the *ganging-up cumulativity* noted by Jäger & Rosenbach (2003, et al.). Here, *Faith* and two bias constraints – viz. *\*self,dis* and *\*pro,co* – are ganging-up against the two other bias constraints.

I will tentatively suggest below that this 'labor sharing' between bias constraints and faithfulness constraints might be related to a second issue which causes some trouble for standard pragmatic accounts of anaphora in the same way the case of Pattern Generalization does, namely the type of differential *self*-marking pattern attested in Dutch.

## 6.5 Languages with multiple reflexivizing strategies

We saw in Chapter 2 how Reinhart & Reuland's reflexivity-based alternative to standard BT was able to account for languages whose reflexive marking patterns discriminate between verbs that are intrinsically reflexive and those that are not.

- (6.54) a. Jan schammt zich  
b. \*Jan schammt zichzelf.  
'John is ashamed.'

- (6.55) a. \*Jan bewondert zich  
b. Jan bewondert zichzelf.

‘John admires himself’

We saw that the fundamental challenge facing R&R’s account was that it is not obviously extendable to other languages. Firstly, reflexive marking is sometimes mandatory for reflexive predicates whether they are intrinsically reflexive or not (cf. English). Secondly, even in languages that bear strong similarities to Dutch, e.g., German or Icelandic, *self*-type marking is not always mandatory for non-intrinsically reflexive predicates. What is more, it is almost certain that it could be shown that the tendency to mark non-intrinsically reflexive predicates is stronger than the tendency to mark intrinsically reflexive ones *across languages*, not just in languages like Dutch where the marking pattern is actually mandatory.<sup>10</sup> R&R’s account would never be able to capture this tendency, since their framework is not equipped to deal with optionality between SELF anaphora and SE anaphora in coargument positions.<sup>11</sup> As we already know, a stochastic, evolutionary OT framework is ideal for capturing such optionality and such tendencies and below I will tentatively suggest how an OT-based evolutionary story might help account for both the strict pattern that we find in Dutch as well as less strict tendencies in other languages.

If this type of pattern could be accounted for in terms of bidirectional learning, it would be a significant step, I think, since it is well known that many languages, not just Dutch, have multiple strategies for expressing reflexive predicates which are to some degree differential. Examples include languages with both verbal and nominal reflexivizing strategies like Russian (Huang, 2000), those with zero versus non-zero reflexivizing strategies like Kannada (Lidz, 1996) and those with single versus multiple emphatics like Turkish (Koenig & Siemund, 2000). Koenig & Siemund (*Ibid.*) have proposed that a cross-linguistic correlation can be recognized between the type

---

<sup>10</sup>Actually, as Reinhart & Reuland (1993, fn. 15) themselves note, the pattern in Dutch is not as strict as it could be. In particular, *zelf*-marked forms are permissible with intrinsically-reflexive verbs but “require discourse justification”. I will ignore exceptions to the paradigm for now, though the fact that they exist only works in my favor, I think.

<sup>11</sup>One can capture optionality in this respect if one is ready to admit that, in certain languages, say German, a SE anaphor can also function as a reflexivizer. But such a move essentially wipes out the distinction R&R draw between SE and SELF anaphora in the first place and leads to circularity. In addition, it would still fail to capture the fact that, cross-linguistically and intralinguistically, *self*-marked objects do exhibit a very strong tendency to be the objects of non-intrinsically reflexive verbs, not intrinsically reflexive ones.



of reflexivizing strategy used and the inherent ‘self-directedness’ or ‘other-directedness’ of the predicate being used.

(6.56) *The predicate meaning/reflexivizing strategy correlation*

The more ‘marked’ a reflexivizing situation (e.g., other-directed), the more marked (i.e., more complex) a reflexivizing strategy will be used to encode it.

It has, on occasion, been claimed that the predicate meaning/reflexivizing strategy correlation can be seen as a direct result of pragmatic implicatures. Huang, for example, states that the correlation is “[c]learly all ... explainable in terms of our M-principle: to convey a marked message, use a marked linguistic expression.” (Huang, 2000, 220) However, while Huang is correct that the M-principle, in spirit, goes well with the marked-form-for-marked-meaning tendency described in the predicate meaning/reflexivizing strategy correlation, an example like (6.55) is not truly explainable in terms of the M-principle; the M-principle says in effect that a marked form  $x$  M-implicates the complement of the I-implicature associated with  $y$  where there is a structural markedness-scale  $\langle x, y \rangle$ . And while there is indeed a structural markedness-scale  $\langle zichzelf, zich \rangle$  – since *zichzelf* is more marked – (6.55b) cannot reasonably be said to M-implicate the I-complement of (6.55a), for, first of all, (6.55a) is not a sentence of Dutch and, secondly, even if it were a sentence of Dutch, it would seem to mean the same as (6.55b).

Differential *zelf*-marking in Dutch, it seems, must be handled as some sort of ‘Pattern Generalization’, and not a case of ‘Antisynonymy’, in the Keenan senses, for just as the reflexive marking of local person pronouns could not be explained in terms of Levinson’s M-principle, Keenan’s Antisynonymy, or Blutner’s notion of weak bidirectional optimality, neither can the discriminatory *zelf*-marking in Dutch; *zich* and *zichzelf* are synonymous.

This case differs from the case of Pattern Generalization that we looked at above, however, since the earlier case involved a marking pattern that applied to potentially ambiguous expressions ‘spreading’ to cases where no ambiguity was present. In the case of *zich* versus *zichzelf* in Dutch, we are presumably dealing with a case where *no* ambiguity was present to begin with in *any* of the cases, for we can assume that an adult hearer would not be at risk to interpret an expression like the ungrammatical *\*Hij bewondert zich* as meaning anything other than ‘He admires himself’. And thus we have every right to wonder why *zelf*-marking showed up in cases like this, since

the speaker is incurring a cost to himself when he could have solicited the correct interpretation without doing so.

I think that if we take for granted that conjoint predicates are more likely intrinsically reflexive than not then we can show how the a marking pattern like the one in Modern Dutch might have shown up as a result of iterated bidirectional learning.

For argument's sake, let us assume an (almost undreamable) scenario wherein a language (I'll call it 'Hypothetical Dutch') *zelf*-marks 10% of all locally conjoint objects, regardless of whether the predicate is intrinsically reflexive or not. Assume too that the intrinsically reflexive predicates outnumber the non-intrinsically reflexive predicates two-to-one and restrict the attention to two inputs.

(6.57) *Frequencies per Hypothetical Dutch*

	Hij schammt SE	Hij schammt SELF	Hij bewondert SE	Hij bewondert SELF
<i>schammt(x, x)</i>	60%	6.7%	0	0
<i>bewondert(x, x)</i>	0	0	30%	3.3%

The major dissimilarity between this experiment and the Keenan's OE experiments discussed above is the fact that *none* of the forms here is potentially ambiguous (assuming that *Faith* is always obeyed). But this does not mean that there must be an absolute lack of hearer-bias in one direction or another, for, as we saw above, bias constraints can either harmonize with *Faith* or clash with it.

In order to capture the fact that a hypothetical learner can recognize some distinction between SE and SELF forms and between the intrinsic and non-intrinsic reflexivity of predicates in order to form biases along these lines, we need to supply him with adequate bias constraints. Let those constraints be as in (6.58), below.

(6.58) *Bias constraints*

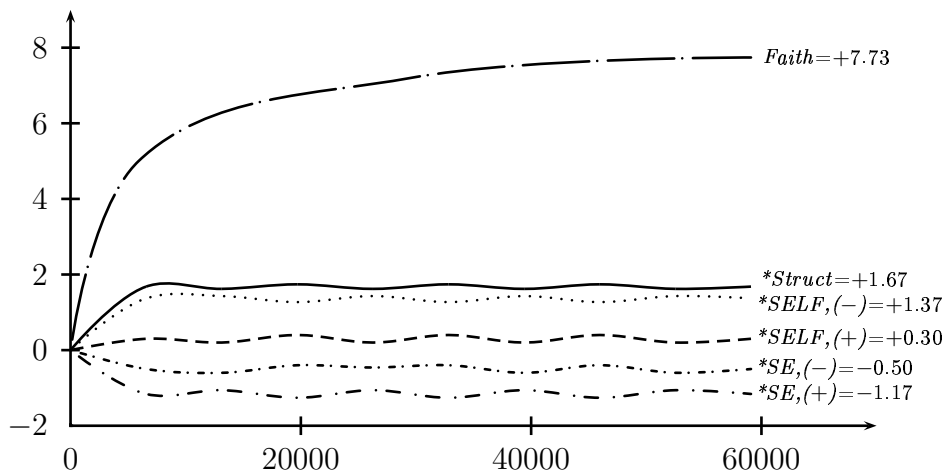
- \**SE, (+)*: SE forms are not objects of intrinsically reflexive verbs.
- \**SE, (-)*: SE forms are not objects of non-intrinsically reflexive verbs.
- \**SELF, (+)*: SELF forms are not objects of intrinsically reflexive verbs.
- \**SELF, (-)*: SELF forms are not objects of non-intrinsically reflexive verbs.

Given the indiscriminate *zelf*-marking in the training corpus, speaker-mode learning will recognize no distinction between intrinsically reflexive and intrinsically non-reflexive predicates and, thus, if speaker-mode learning were the only consideration,  $*SE,(+)$  and  $*SE,(-)$  would be ranked about equally, as would  $*SELF,(+)$  and  $*SELF,(-)$ .

However, hearer-mode learning will recognize an asymmetry in the number of intrinsically reflexive and intrinsically non-reflexive predicates, since the learner will be exposed to twice as many of the former as compared to the latter. For that reason, both SE forms and SELF forms will be learned as more likely associated with *schammen* as opposed to *bewonderen*. In other words,  $*SELF,(-)$  will outrank  $*SELF,(+)$  and  $*SE,(-)$  will outrank  $*SE,(+)$ .

A BiGLA-learned grammar based on the frequencies in (6.57) looked as below.

(6.59) *Bidirectional learning, per (6.57) (first generation)*



Note that because non-intrinsically reflexive inputs are the rarer type of inputs, hearer-mode learning has ranked i.e.,  $*SE,(-)$  and  $*SELF,(-)$  higher than their respective counterparts,  $*SE,(+)$  and  $*SELF,(+)$ . As a result, the former two have been learned closer to  $*Struct$ . In this way, hearer-mode learning will induce the same sort of ‘accidental’ generative effects we saw before; we will again see a case in which the marked form is ‘repelled’ from the statistically more prevalent meaning (though these effects are more

subtle this time, since there is only a 2-to-1 asymmetry in the training corpus between the two types of inputs).

The actual output frequencies for the grammar in (6.59) looked as follows.

(6.60) *Bidirectional learning (first generation)*

	Hij schammt SE	Hij schammt SELF	Hij bewondert SE	Hij bewondert SELF
<i>schammt(x, x)</i>	60.5%	6.2%	0	0
<i>bewondert(x, x)</i>	0	0	29.7%	3.6%

In this case, however, although hearer-mode has learned the bias constraints in a way that reflect hearer-bias to the effect that SE and SELF forms are usually associated with intrinsically reflexive predicates, this bias will not usually affect the final interpretation of a sentence like *John bewondert zich(zelf)*, since we are assuming that *Faith* will guarantee that an individual with the grammar in (6.59) will not (typically) interpret *bewonderen* as meaning *schammen*. In this way, the hearer-mode bias will be ‘overruled’ by *Faith*.

Note that in the learned grammar in (6.59) *Faith* must be ranked extremely high to ensure that it will overrule the bias constraints. The exact trends in learning effects and the mechanics of the relationships between the constraints in an experiment like this will have to remain an area of further research, but at least one thing is clear: we know that *Faith* is only ranked high because the learner has drawn many false hypotheses during the learning process (especially very early on). If he had never drawn the wrong hypothesis, *Faith* would be at zero. We must infer that the incorrect hypotheses he drew were drawn because of the bias constraints. (*Faith* doesn’t compete with anything else.) Moreover, from the looks of the learning curves, *Faith* is still going up.<sup>12</sup>

But if the marking pattern were categorically differential with respect to the ‘intrinsically reflexive/non-intrinsically reflexive’ dimension, *Faith* would not need to overrule the bias constraints, or at least not nearly as often. And if this were true, *Faith* would need to be promoted much less, since the bias constraints and *Faith* would be working in the same direction.

<sup>12</sup>If one runs the same experiment with 600,000 inputs instead of 60,000, *Faith* commonly goes up to +10 or higher, while the other constraints remain roughly as above.

To appreciate how such a marking pattern could reinforce faithfulness in this case, just consider again a simple thought experiment wherein we have a newborn learner whose constraints are set at zero. Suppose he observes his first input: *Hij bewondert zichzelf*, meaning ‘He admires himself’.

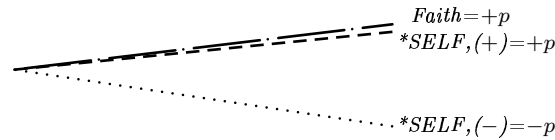
Suppose this learner draws the correct hypothesis in speaker mode, but that his hypothesis in hearer-mode is wrong. Since we are only considering two inputs, his hypothesis must be ‘He is ashamed’ (i.e., the meaning of ‘*Hij schaamt zich*’). The relevant learning effects are:

(6.61) *Hearer-mode learning effects per ‘Hij bewondert zichzelf’*

Observed pair	hypothesis	<i>Faith</i>	<i>*SE,(-)</i>	<i>*SE,(+)</i>	<i>*SELF,(+)</i>	<i>*SELF,(-)</i>
<i>Hij bewondert zichzelf</i> ‘He admires himself.’	‘He is ashamed.’	↑			↑	↓

The adjustment would effect the scenario in (6.62), below.

(6.62) *(Partial) rankings after first observation*



Now suppose that the learner observes a second occurrence of *Hij bewondert zichzelf*, with the usual meaning.

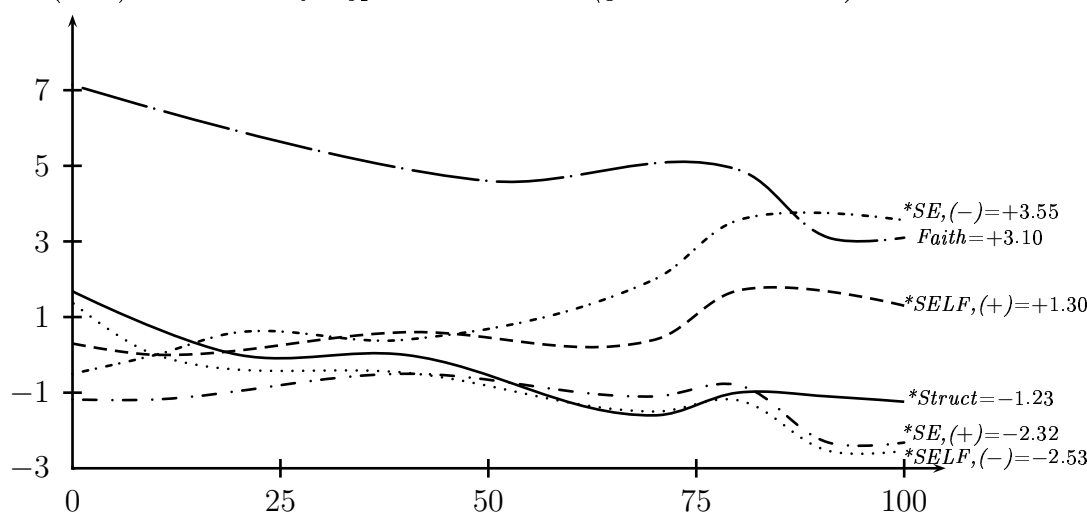
This learner is now less likely to interpret *Hij bewondert zichzelf* unfaithfully as ‘He is ashamed’ than he was when his constraints were ranked at zero. Not only because *Faith* has been ranked higher, but also because the comparatively higher ranking of *\*SELF,(+)* (and the comparatively lower ranking of *\*SELF,(-)*) would direct him not to interpret a sentence like *He bewondert zichzelf* as meaning ‘He is ashamed’, or as meaning anything else that involved an intrinsically reflexive predicate. Now, if the marking pattern is non-differential (or not very differential) then a learner will later observe

many occurrences of *Hij bewondert zich* and *Hij schaamt zichzelf*, and the faith/bias alliance will not be preserved. But if the pattern is consistent, the alliance just gets reinforced.

Because I have restricted my attention to only two verbs, the effects are a bit overblown in the thought experiment. However, the point remains that, where a marking strategy is differential, bias constraints and faithfulness constraints will generally favor the same interpretations for the same forms. Where there is less conflict between *Faith* and the bias constraints, a learner is then less likely to draw an incorrect hypothesis. The less incorrect hypotheses a learner draws, the more stable his grammar is. In this way, the cooperation of bias constraints and faithfulness constraints increases the stability of a grammar.

A grammar like the one (6.59) commonly stabilized as a categorical split marking pattern.<sup>13</sup> After 100 generations of iterated bidirectional learning, our Hypothetical Dutch took the evolutionary path shown below.

(6.63) *Evolution of Hypothetical Dutch (generations 0-100)*



<sup>13</sup>Admittedly though, this result was not nearly as common as the results shown in the previous experiments in this chapter, which were repeatable 100% of the time. However, in my experience, they occur at levels much better than chance and are even more likely when the common-input/rare-input ratio is more imbalanced. (Recall that here it is only 2-to-1.) Cf., the ‘Hypothetical Greek’ example in the next section for an illustration of this.

One can see that the first 20 generations of Hypothetical Dutch are somewhat reminiscent of the evolution of Keenan's OE. Here, *\*SELF,(+)* and *SELF,(−)* – i.e., the bias constraints relevant to the interpretation of marked forms (analogous to *\*self,dis* and *\*self,co* in the previous experiment) – have undergone cross-generational adjustment due to the accidental generative effects of hearer-mode learning. Because these accidental effects result in more SELF outputs for non-intrinsically reflexive predicates and less SELF outputs for intrinsically reflexive ones, hearer-mode learning of subsequent generations will learn *\*SELF,(+)* higher and *SELF,(−)* lower.

Note too the significant cross-generational demotion of *Faith*. Remember that this does not mean that *Faith* ever got demoted. It has simply been promoted less with passing generations. The reason for this is again a ‘conspiracy’ of sorts between the bias constraints and *Faith*. The more differential the marking pattern is with respect to the intrinsic/non-intrinsic reflexivity dimension, the less chance there is that a learner's hypothesis will be unfaithful, since he will not only be able to rely on *Faith* to hypothesize correctly, but also on bias constraints.

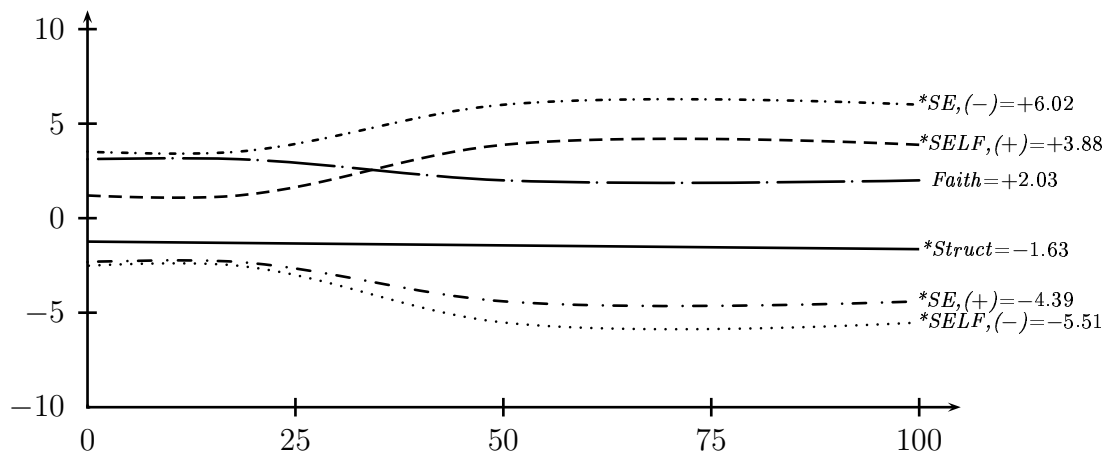
The actual output frequencies for the evolved grammar in (6.63) were as below.

(6.64) *Frequencies per (6.63)*

	schammt	He schammt	He bewondert	He bewondert
	SE	SELF	SE	SELF
<i>schammt(x, x)</i>	62%	4.7%	0	0
<i>bewondert(x, x)</i>	0	0	.9%	32.4%

Using the frequencies in (6.64) as an ancestor corpus and executing 100 additional generations of iterated learning evolution allows us to see the grammar above stabilize completely.

(6.65) *Evolution of Hypothetical Dutch (generations 100-200)*



The differential marking pattern has now grammaticalized entirely:

(6.66) *Frequencies per (6.65)*

	schammt SE	He schammt SELF	He bewondert SE	He bewondert SELF
<i>schammt(x, x)</i>	66.7%	0	0	0
<i>bewondert(x, x)</i>	0	0	0	33.3%

I have tried to illustrate with the experiment above why we might have reason to believe that the stability of a bidirectionally learned grammar can depend partly on a lack of dissonance between faithfulness constraints and bias constraints. If this were true, grammars would tend to evolve in a way in which faithfulness constraints and bias constraints came to share the labor of optimization rather than conflict with each other. On this picture, the differentiation that takes place is due to a harmonization of sorts between *Faith* and the bias constraints, not blocking or ‘Antisynonymy’.<sup>14</sup>

<sup>14</sup>Note that though bidirectional optimization is virtually irrelevant in terms of the evaluations that determine the actual outputs for each generation, this does not mean that bidirectional optimization does not play a role in the evolution of the grammar. In particular, bidirectional optimization can potentially play a role in determining a correct *hypothesis* in speaker mode. This can be very relevant, especially early in the learning process when *Faith* is closely ranked with the bias constraints. In fact, evolutionary experiments like the one above run *without* the effects of bidirectional optimization were



If this explanation is even in a broad sense correct, it would seem to be applicable to other multiple reflexivizing strategies that follow the predicate meaning/reflexivizing strategy correlation noted by (Koenig & Siemund, 2000) and mentioned above. Furthermore, it might be in principle extendable to cases of double-marking in languages Japanese, Padovano and Spanish, discussed in Chapter 2, which generally trouble the analysis of Reinhart & Reuland (1993, et al.) as well as the account of Levinson (2000), whose GCI-based picture is difficult to reconcile with marking that is not directly related the solicitation of some M-implicature. However, I will leave the details in this regard as an open question for now.

## 6.6 Loose Ends

There are still quite a few of issues that the experiments conducted so far do not address directly and I will briefly say below why I believe that some of what has been said above might, in principle, be extendable to other aspects of binding phenomena. In particular, issues related to c-command and the Thematic Hierarchy Condition, LDAs, and the distribution of R-expressions might benefit from an approach like the one discussed in the previous two sections.

### 6.6.1 C-command & the Thematic Hierarchy Condition

One blatant challenge that meets standard binding theories are languages with nominative anaphors such as Hungarian, Malagasy, and Greek, discussed in Chapter 2.

We saw how Jackendoff (1972), Grimshaw (1990), et al. have addressed the problem in terms of a Thematic Hierarchy Condition whereby a reflexive may not outrank its antecedent on some thematic hierarchy. Ntelitheos (2001) has already shown how the commonsense strategy of stating the THC as an OT constraint and letting it be ranked among constraints that represent Principle A and B-type preferences can account for the type of cross-linguistic variation between, say, Greek and English. Very roughly: where Principle A outranks the THC, nominative anaphors are ungrammatical, otherwise not.

---

less likely to evolve in the way that Hypothetical Dutch did. However, I must leave the details of how much probabilistic influence bidirectional optimization has at various stages in the learning process as a point for further inquiry.

However, an evolutionary account based on iterated bidirectional learning might offer even more promise in terms of explaining the THC-like effect exhibited in Greek and other languages.

To illustrate how, we can assume that a learner can establish biases with respect to configurational relations and thematically hierarchical relations. For example, we could once again restrict our attention to simple transitive clauses and a set of two inputs and two outputs. The inputs:  $\{S >_{\theta} O, O >_{\theta} S\}$ , i.e., one input such that the subject is higher than the object on the Thematic Hierarchy, and a second such that the opposite is true. The outputs:  $\{[NP... \alpha], [\alpha... NP]\}$ , i.e., a syntactic structure wherein the antecedent c-commands the anaphor and one wherein the anaphor c-commands the antecedent.

Again, we can form bias constraints in the usual way, i.e., conjunctions representing dispreferences for each input-output pair-type, per the two relevant dimensions.

(6.67) *Bias constraints*

$*[NP... \alpha], S >_{\theta} O$ : Antecedents do not precede reflexives where the subject is higher than the object on the Thematic Hierarchy.

$*[NP... \alpha], O >_{\theta} S$ : Antecedents do not precede reflexives where the subject is lower than the object on the Thematic Hierarchy.

$*[\alpha... NP], S >_{\theta} O$ : Reflexives do not precede antecedents where the subject is higher than the object on the Thematic Hierarchy.

$*[\alpha... NP], O >_{\theta} S$ : Reflexives do not precede antecedents where the subject is lower than the object on the Thematic Hierarchy.

We have seen above how a set of bias constraints alone effectively makes it possible to learn virtually any grammar stably and would make little in the way of evolutionary predictions. However, we could surmise that the four bias constraints above might bear a relationship to some universal markedness constraint, say one related to linear prominence and canonical role.

(6.68) *Canon*: Canonical subjects are left of objects, canonical objects are left of oblique arguments.

We can imagine ‘Hypothetical Greek’, wherein the majority of inputs, say 90%, are  $S >_{\theta} O$ , and 10% of all inputs are associated with ‘marked’

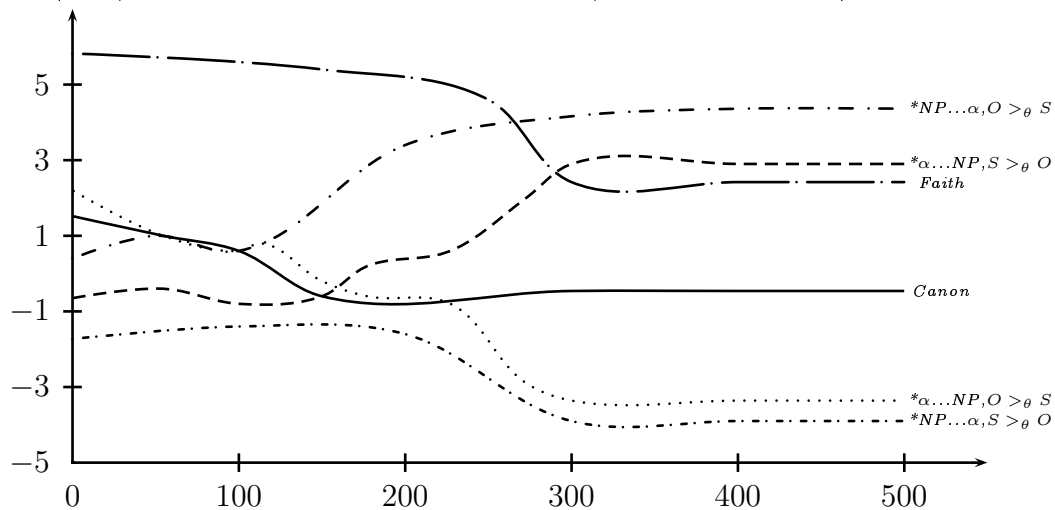
outputs, where in this case ‘marked’ means constructions which violate the constraint *Canon*.

(6.69) *Hypothetical Greek*

	Petro hit himself	Himself hit Petro	Petro pleases himself	Himself pleases Petro
<i>hit(p, p)</i>	81%	9%	0	0
<i>pleases(p, p)</i>	0	0	9%	1%

These numbers may or may not be realistic, but they serve well enough to illustrate again the point made earlier in the Hypothetical Dutch experiment. (This experiment is exactly the same except that instead of the 2-to-1 ratio that I assumed to hold among intrinsically reflexive versus non-intrinsically reflexive predicates, I assume in this case that the more common inputs outnumber the rarer ones 10-to-1.) In a great majority of BiGLA experiments, Hypothetical Greek stabilizes as we would expect. For example, an experiment which simulated 500 generations of iterated learning using (6.69) as an ancestor corpus looked roughly as in (6.70), below.

(6.70) *Evolution of Hypothetical Greek (generations 0-500)*



This is just another case in which the stability of the grammar increases as conflict between *Faith* and the bias constraints subsides.

Now consider again the relevant data from Modern Greek, discussed in Chapter 2 and repeated below.

(6.71) Modern Greek (Everaert & Anagnostopoulou, 1997)

a. \*O eaftos tu<sub>i</sub> ton xtipise ton Petru<sub>i</sub>.

‘Himself hit Petro.’

b. O eaftos tu<sub>i</sub> tu aresi tu Petru<sub>i</sub>.

‘Himself pleases Petro.’

Per the evolved grammar in (6.70), (6.71a) is out, since  $*[\alpha...NP], S >_{\theta} O$  greatly outranks  $*[NP... \alpha], S >_{\theta} O$ . Meanwhile, (6.71b) is fine, since  $*[NP... \alpha], O >_{\theta} S$  greatly outranks  $*[\alpha...NP], O >_{\theta} S$ . In this way, we can view the THC-like pattern as case of stabilization through differentiation and another manifestation of the marked-forms-(only)-for-marked-meanings pattern, deriving it by way of bidirectional learning and the interaction of bias constraints with a universal markedness constraint like *Canon* and a universal faithfulness constraint like *Faith*.

The actual frequencies for the evolved grammar in (6.70) were:

(6.72) *Evolved Hypothetical Greek (500 generations)*

	Petro hit himself	Himself hit Petro	Petro pleases himself	Himself pleases Petro
<i>hit(p, p)</i>	89.81%	.19%	0	0
<i>pleases(p, p)</i>	0	0	.01%	9.9%

Of course, in this example, as in the others, restricting the attention to such a small set of inputs, outputs, and constraints makes such results easier to come by, and a full description of how certain dimensions of bias can be very relevant in some languages but not in others will have to remain an object of further research.

### 6.6.2 LDAs

A second gap in the discussion above is the issue of long-distance Anaphors (LDAs), discussed in Chapters 2 and 3. The only experiments conducted above were those which involved simple transitive clauses, so LDAs were a moot point in that discussion. However, I think the framework used above could prove helpful in modelling the evolution of LDAs and the tendencies they tend to follow both across and within languages.

Huang (Huang, 1994, 2000, et al.) has long argued that the distributional behavior and resolution of LDAs cannot be explained in purely syntactic terms, especially for languages like Chinese, Japanese, and Korean (Huang calls these ‘*pragmatic languages*’), which often break well known ‘rules’ for LDAs, e.g., their supposedly universal subject-orientation or morphological simplicity. (Some LDAs can even appear unbound.) To partly deal with the problem of tendencies-but-not-universals that shows up so often with LDAs, Huang has proposed a pragmatic resolution strategy meant to reflect some of the morphological and interpretational tendencies of LDAs, the ‘antecedent search procedure for reflexives’ (Huang, 1994, 178).

(6.73) *Huang’s antecedent search procedure for reflexives*

In a structure sort  $[s_1[s_2 R]]$ , where  $R$  is a reflexive,  $R$  is interpreted as referentially dependent according to the following preference order:

1.  $R$  is referentially dependent on the local subject; failing which:
2.  $R$ , especially when morphologically complex, is referentially dependent on the local object; failing which:
3.  $R$ , especially when morphologically complex, is referentially dependent on both the local subject and the object (split antecedents); failing which:
4. 1-3 is recursively applies to the next, higher clause, until an antecedent is found; failing which:
5. find the nearest antecedent in the discourse, preferably a topic; failing which:
6. settle for an ‘arbitrary’ interpretation

In the restricting our attention to SE/SELF type reflexives for the moment, Huang’s 1-3 can effectively be restated in OT as a ranked set of bias constraints (Constraint 1 is the highest here, 6 the lowest):

(6.74) *Bias-constraint-based restatement of (6.73) (steps 1-3)*<sup>15</sup>

1. *\*SE,LocalSubject&Object*
2. *\*SELF,LocalSubject&Object*
3. *\*SE,LocalObject*
4. *\*SELF,LocalObject*
5. *\*SE,LocalSubject*
6. *\*SELF,LocalSubject*

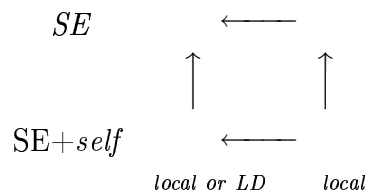
Moreover, there might be no need to stipulate any universal ranking for the constraints above. Rather, the ranking in (6.74) might just be *derivable* in a context of evolutionary bidirectional learning. In particular, this case is similar to the cases Hypothetical Dutch and Hypothetical Greek where (a) ‘accidental’ generative consequences can result from hearer-mode learning of the bias constraints and the relationship between those constraints and structural markedness constraints; marked forms tend to gravitate toward marked meanings, and (b) if the marking becomes partially differential (as a result of (a)), the grammar tends to stabilize in a way such that the dissonance between bias constraints and *Faith* is reduced, reinforcing the differentiation.

I think an evolutionary analysis based on bias and iterated bidirectional learning might help offer an answer to the question of why morphologically simplex anaphora tend to be less restricted than those which are morphologically complex. In short: because bidirectional learning of bias constraints, markedness constraints, and faithfulness constraints has the effect of causing structurally marked expressions to gravitate toward rare inputs, it will basically follow from this that unmarked things will be less restricted things

---

<sup>15</sup>I forgo representing the recursive step 4 or steps 5 and 6 only for the purpose of brevity. I have embellished a little here too, since I have represented distinctions between complex and simplex anaphora in one place where Huang does not. Such an addition might actually be an accurate one, though it is not crucial to the point here.

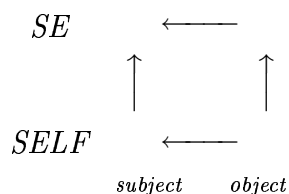
(6.75)



Moreover, it will also follow that where structurally unmarked forms *are* restricted, they are restricted to the *more common* range of cases, not the rare range.

For example, if we take for granted the fact that canonical subjects are statistically more prevalent in language use than objects are, we could fully expect to see attested patterns of differential behavior on these dimensions of the same sort that was seen in, e.g., the experiment involving Hypothetical Dutch, which showed how multiple strategies might tend to be differential because such patterns are more faithful and more stable.

(6.76)



If such a line could be properly spelled out, we might be on our way to an account that explained differences in languages like Norwegian, where such a pattern may have grammaticalized (since *seg (selv)* is subject-oriented and *ham selv* object-oriented (Hellan, 1988)) as well as languages wherein the pattern was a tendency, not an exceptionless phenomenon.

Such an analysis might also be extended to account for other tendencies with respect to LDAs, e.g., their tendency to resist split antecedents (Saxena (1985), Davison (1999)) and perhaps their tendency to take sloppy (bound variable) not strict readings (falsely claimed to be a universal by Williams (1977), Bouchard (1984), and Lebeaux (1985)).

(6.77) Icelandic (Thráinsson, 1991)

\*Jón<sub>i</sub> sagði Maríu<sub>j</sub> að þú hefðir svikið sig<sub>i+j</sub>.  
 John told Mary that you had betrayed SE.'  
 'John told Mary that you betrayed them.'

(6.78) Hindi/Urdu (Davison, 1999)

Gautam<sub>i</sub> apnee<sub>i</sub>-koo caalaak samajhtaa hai aur Vikram  
 Gautam SE-DAT smart consider-IMPF is and Vikram  
 bhii.  
 too  
 'Gautam considers himself smart, so does Vikram.'

We might take these facts to be an evolutionary consequence of bidirectional learning, plus the fact that split antecedents are much less statistically common than non-split ones, and sloppy uses of anaphora are more prevalent than strict coreferential ones.

(6.79)



However, to what extent these tendencies tend to grammaticalize will depend greatly on what type of ratios exist between the various types of inputs mentioned here, among other things, and those details must remain an area of further research for now.

In addition to the issues above, an evolutionary learning account of LDA patterns might also prove suitable for capturing language specific configurational sensitivities that LDAs sometimes exhibit, as well as the semantic differences they can bear to pronouns.

First of all, as noted in Chapter 3, Levinson (2000), et al. have claimed that LDAs are “always logophoric” and thus an LDA like Icelandic *sig*, it is



claimed, solicits a shift in point of view or perspective (rather than a reflexive reading) where a pronoun could induce no such interpretation. Levinson suggests that his Stage 1-3 account reflexive marking is applicable to logophoric marking strategies as well and, indeed, historical data suggests that the marking of logophoricity by means of a SE-type expression or the like is a gradually evolving phenomena.<sup>16</sup>

Without argument, I think that with enough bias constraints, e.g., *\*pro*, *+log*, *\*SE,co*, and the like, we could model differentiation between regular pronouns and SE anaphora based on the semantic distinction of logophoricity and traditional bidirectional optimization. Where SE is the marked form (since it violates  $\varphi$ , cf. Chapter 4) it pairs with the marked, logophoric reading.

On the other hand, Burzio's syntactically-based account of LDAs easily captures cases where SE anaphora do apparently follow some configurational guidelines; one important class of cases are those LDAs which are not only *allowed* to appear long-distance in a certain type of embedded clause, but are *mandatory* in those clause-types since pronouns are ungrammatical.

(6.80) Icelandic (Maling, 1984, 212)

Haraldur<sub>i</sub> skipaði mér að raka sig<sub>i</sub>/*\*hann<sub>i</sub>*.  
 Harold asked me that shave-INF SE/him  
 'Harold asked me shave him'

Levinson's remarks on LDAs to the effect that they are always logophoric and always bear some semantic difference compared to pronouns would not lead us to expect that the optionality between LDAs and pronouns would ever break down in its way, and thus it appears that we again need an explanation of how the restriction grammaticalized.

Again without argument, I think that something like Burzio's constraints could be integrated into an evolutionary bidirectional learning experiment to provide the precision necessary to capture these cases. For instance, if we were to introduce bias constraints into the analysis such as *\*[NP<sub>i</sub> ... [<sub>Indic</sub> NP<sub>j</sub>...SE]]* and *\*[NP<sub>i</sub> ... [<sub>Sbjv</sub> NP<sub>j</sub>...*pro*]]*, and so on, we could reflect the appropriate distinctions between various clause-types in our experiments.

---

<sup>16</sup>According to Magnusson (1985), e.g., LDAs were much rarer in Old Icelandic than they are in Modern Icelandic.

Moreover, I think we could hope that rather than stipulating the universal ranking for the constraints as Burzio does with his Optimal Antecedent Hierarchy (cf. Chapter 2), we could *derive* the universal ranking through the consideration of extralinguistic facts and their effects on training corpora.

For example, it seems almost ridiculous *not* to believe that in, say, Old Icelandic, embedded subjunctive clauses were statistically more likely to contain logophoric triggers than embedded indicatives.<sup>17</sup>

If it were the case that SE forms had become or were becoming more strongly associated with logophoric meanings then this would have an effect on the promotion and demotion of bias constraints like  $*[NP_i \dots [IndicNP_j \dots SE]]$  and  $*[NP_i \dots [SbjvNP_j \dots SE]]$ ; for the former might be learned as outranking the latter since embedded indicatives might, statistically, contain far less LDAs than embedded subjunctives. This would lead to another pragmatic-to-structural story, where a marking device meant to convey logophoricity came to be associated with a syntactic property after  $n$  generations of iterated bidirectional learning.

I do not know whether we can truly find that statistical correlations could get us Burzio's Optimal Antecedent Hierarchy for free, but the idea that, say, infinitives are a more common source of logophoric triggers than indicatives, and small clauses even more so than infinitives, seems perfectly reasonable to me. Such an approach might open up a way for giving a diachronic story of, say, Old-to-Modern Icelandic that explained how the configurational sensitivities had grammaticalized. Moreover, it would allow us to leave room for the integration of discourse-related and thematic-related constraints in addition to syntactic ones, and thus we could hope for a unified account of 'syntactic languages' and 'pragmatic languages', in Huang's terms.

I must leave specific experimental support for all of this as yet another area of further research, though.

### 6.6.3 Principle C effects

One final point I shall mention is Principle C and its supposed effects. We saw in Chapter 2 that Chomsky's Principle C has been, for the most part, given up on in generative linguistics, largely due to the fair number of counterexamples one is able to find in virtually any language.

---

<sup>17</sup>I choose Old Icelandic again since, according to Sigurðsson (1990), the *sig* of Old Icelandic did *not* exhibit configurational discrimination of the kind manifested in Modern Icelandic.

To my knowledge, no one has ever attempted to spell out explicitly what kind of discourse conditions ‘warrant’ the repetition of R-expressions intrasententially; in most counterexamples that one finds it is fair to say that the repetition of the R-expression is related to some sort of emphasis, humor, or both.

One formal attempt to improve on Principle C was that of Lasnik (1989), who divided it into two conditions, one parameterized and one not, as to predict that R-expressions can be bound by other R-expressions, but not by pronouns. I have already discussed why Lasnik’s  $C_1$  and  $C_2$  are both too strong and too weak for their intended purposes, however, I think that the general spirit of both Chomsky’s Principle C and Lasnik’s  $C_1$  and  $C_2$  could be captured in a stochastic OT framework to get to the correct prediction: viz. that R-expressions are typically unbound, and where R-expressions do appear bound they will tend to be bound by other R-expressions, not by pronominals.

As for the first point, we can ask ourselves what type of markedness constraint is at work that would generally induce an avoidance of R-expressions. Levinson (1987b, 1991), Burzio (1989, 1998), Huang (1994, 2000), Richards (1997), and others have all suggested a notion of *referential economy*, whereby, basically, R-expressions are ‘fully-referential’ and least economical, reflexives are non-referential and thus most economical, and pronouns fall somewhere in between.

I am personally very suspicious of notions such as ‘referential economy’, however, if one wanted to adopt a universal markedness constraint in order to reflect referential economy, one could. In doing so, I think that one could also easily build an analysis of why grammars evolve in such a way that the marked, referentially uneconomical form is not employed when a pronoun or reflexive is available.

On the other hand, I do not think that an evolutionary account of Principle C necessarily requires any reference to the notion of referential economy, but rather might be a manifestation of purely structural/morphological economy considerations. The obvious argument *against* such an idea has been given by Levinson (1987a) who argues that the low costs of using anaphoric pronouns cannot be directly related to articulatory costs, since their employment does not always result in a reduction of articulatory effort (compare Chinese *ziji* with a name like *Mao*, for example).

However, I think that the mechanics of the evolutionary story told above are not challenged by this objection. Specifically, we could imagine a speech

community wherein speakers used an anaphor if and only if the use of an anaphoric expression actually did result in the reduction of articulatory effort. Let us say, for argument's sake, that just 60% of the names, definites, indefinites and so on were such that the cost of producing them was higher than the cost of producing a pronoun. If we assumed that a learner growing up in such a community had access to a set of bias constraints, perhaps *\*R-expression,bound*, *\*R-expression,unbound*, *\*pro,unbound*, and so on, then we know that he will rank the bias constraints according to the learning data and, just like in the examples of Zeevat & Jäger (2002), Cable (2002), and my own analysis above, the learner will have a grammar that demands 'structural marking' (or in this case, structural pronoun usage) where the teacher's grammar made no such demand. I.e., 'pragmatic' usage of pronouns done strictly for reasons of articulatory economy has now become a structural usage. And the evolution of that pattern could proceed as expected. In this way, one can avoid Levinson's objection while still avoiding reference to mysterious notions of 'referential economy' (whereby reflexives like *himself* are taken to be more economical than simple pronouns like *him*).

As for Lasnik's observation that R-expressions discriminate with respect to whether they can be bound by pronouns or other R-expressions, I think that such a phenomenon might also be explainable in terms of a bias-based evolutionary learning story. For example, if we take for granted that it is a universal of language use that unreduced NPs are *more likely* to be used as antecedents than anaphora and that the opposite is true of reduced NPs (perhaps for reasons related to information processing, as Hawkins (2002) has suggested), then we could imagine this bias grammaticalizing after *n* generations of iterated bidirectional learning in the same way that biases have become grammaticalized in the experiments above.

Again, though, I will leave these points as matters of speculation for now.

# Chapter 7

## Conclusion

Above I have argued for a pragmatic, evolutionary account of basic binding phenomena based primarily on the notions of bidirectional optimization and bidirectional learning.

The original motivation for giving such an account was to map a clear evolutionary path between anaphoric patterns that are ‘pragmatic’ and those which are ‘structural’, where a ‘pragmatic’ pattern is an optional marking pattern dependent on contextual factors, discourse factors, and so on, and a ‘structural pattern’ is one that is not immediately influenced in this way.

Primitives like Chomsky’s  $\pm$ Pronominal and  $\pm$ Anaphoric were thrown out, as were Reinhart & Reuland’s notion of  $\pm$ Reflexivizing function,  $\pm$ Referential independence, Levinson’s notion of  $\pm$ Logophoric, and the like.

For this reason, categorical restrictions – such Principles A and B, the Thematic Hierarchy Condition, and even defeasible pragmatic stipulations like the Disjoint Reference Presumption were discarded as well.

Instead, properties like ‘+Reflexivizer’, ‘–Local’, ‘+Logophoric’, and so on, were all *relativized* and *probabilized* by way of *bias constraints* in a stochastic OT framework. The extent to which a form  $f$  has a syntactic/functional property  $P$  is now a function of  $n - m$ , where  $n$  and  $m$  are the ranking values of two bias constraints,  $*f, P$  and  $*f, \neg P$ . Any ‘hard rules’ in these regards that show up in a grammar are seen as products of evolutionary learning and, at base, it is assumed that these bias constraints could be learned in any way imaginable.

However, I have tried to show above why basic binding phenomena – especially those which follow the *marked-forms-for-marked-meanings* pattern, noted by Atlas & Levinson, Horn, Blutner, and others – follow (probabilis-

tically) from the interaction of bias constraints with two other types of constraints, *markedness* constraints and *faithfulness* constraints in a context of bidirectional learning.

In summary, the reasoning behind that story went as follows:

1. Wherever there is a statistical asymmetry between two types of inputs/meanings, one rarer and one more common, hearer-mode learning will always rank bias constraints that militate against rare inputs/interpretations higher than those which militate against common ones, all things being equal.
2. Since the set of bias constraints is assumed to be comprehensive, it follows that, for any universal markedness constraint  $C$ ,  $C$  will ‘gang-up’ *with* a certain subset of bias constraints and gang-up *against* certain others.
3. Where universal markedness constraint  $C$  must be learned with a set of bias constraints, the grammar will (probabilistically) converge in such a way that  $C$  is ranked as to accurately reflect the generative optimization for more *common* inputs, rather than rare ones (if it cannot do both).
4. It follows (probabilistically) from 1 through 3 that where there is dissonance between hearer-mode and speaker-mode learning, marked forms will become more tolerable for rare inputs, simply as an ‘accidental’ consequence of hearer-mode learning.
5. The resulting imbalance will induce future generations of learners to learn constraint rankings that reflect that imbalance.
6. This effect is ‘self-reinforcing’, since each future generation will see a greater imbalance than the previous one.
7. Moreover, differential marking patterns that are not categorically differential tend to become categorically differential either (a) due the effects of bidirectional optimization (since categorical patterns reduce ambiguity) or (b) due to a resolution of conflict between bias constraints and faithfulness constraints (since categorical marking patterns increase faithfulness and learnability.)

On such an account, one can easily derive what Horn called the ‘division of pragmatic labor’ – as well as certain pragmatic stipulations relied upon in the literature, such as the Disjoint Reference Presumption – and give a precise account of how such ‘divisions of labor’, ‘presumptions’, ‘stereotypes’, and so on, actually manifest themselves in grammatical knowledge, thus giving a formal explanation for what a pragmatic ‘fossil’ really is.

How far such an approach can actually get when is tested against more real data remains to be seen. However, the initial results seem promising and the approach might provide a powerful way of offering formal justification to many other areas of pragmatics and the pragmatics/syntax interface.

I will tout the merits of the approach no further, however, for – to cite a beautiful Haitian proverb on the importance of humility as well as an equally beautiful example of a fully grammaticalized reflexive... *Sel pa vante tèt li di li sale.* (‘Salt doesn’t brag that it’s salted.’)

# Bibliography

- Aikawa, T. (1993). *Reflexivity in Japanese and LF Analysis of Zibun-Binding*. PhD thesis, Ohio State University.
- Aissen, J. (1999). Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory*, 17, 637–711.
- Aissen, J. (2000). Differential object marking: Iconicity vs. economy. ms., University of California, Santa Cruz.
- Aissen, J. (2003). Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21, 435–483.
- Atlas, J. D. & Levinson, S. (1981). It-clefts, informativeness and logical form: radical pragmatics. In P. Cole (Ed.), *Radical Pragmatics* (pp. 1–61). London: Academic Press.
- Austin, P. (1981). Case marking in southern pilbara languages. *Australian Journal of Linguistics*, 1, 211–226.
- Austin, P. (1987). Cases and clauses in jiwari, western australia. Paper presented at the University of California, Los Angeles.
- Bar-Hillel, Y. & Carnap, R. (1952). An outline of a theory of semantic information. Technical Report 247, MIT. Reprinted in Y. Bar-Hillel, 1964, *Language and information* pp.221-274. Reading, MA, Addison-Wesley.
- Battistella, E. (1989). Chinese reflexivization: A movement to infl approach. *Linguistics*, 27, 987–1012.
- Battistella, E. & Xu, Y. (1990). Remarks on the reflexives in chinese. *Linguistics*, 28, 205–240.



- Beaver, D. & Lee, H. (2003). Form-meaning asymmetries and bidirectional optimization. In J. Spenader, A. Eriksson, & Östen Dahl (Eds.), *Variation within Optimality Theory* (pp. 138–148). Stockholm: University of Stockholm.
- Benedicto, E. (1991). Latin long-distance anaphora. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 171–184). Cambridge: Cambridge University Press.
- Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge: MIT Press.
- Blutner, R. (2000a). Bidirectional optimization in natural language interpretation. Presented at the Utrecht Conference on the Optimization of Interpretation, Utrecht; <http://www.blutner.de/utrecht.pdf>.
- Blutner, R. (2000b). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17(3), 189–216.
- Blutner, R. (2001). Optimality theory and natural language interpretation. Presented at the 13th Amsterdam Colloquium, Amsterdam; <http://www.blutner.de/ac.pdf>.
- Blutner, R. (2002). Optimality theory and grammaticalization. Presented at the Workshop on the Roots of Pragmasemantics (with Hanneke van der Grinten), Szklarska Poreba, Poland; <http://www.blutner.de/sklarska2002.pdf>.
- Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD thesis, University of Amsterdam.
- Boersma, P. & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32, 45–86.
- Bouchard, D. (1984). *On the Content of Empty Categories*. Dordrecht: Foris.
- Bresnan, J. (1998). Optimal syntax. In J. Dekkers, F. van der Leeuw, & J. van de Weijer (Eds.), *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford University Press.

- Bresnan, J. & Kaplan, R. (1982). Lexical functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations* (pp. 173–282). Cambridge, MA: MIT Press.
- Briscoe, T. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2), 245–296.
- Buchwald, A., Schwartz, O., Seidl, A., & Smolensky, P. (2002). Recoverability optimality theory: discourse anaphora in a bidirectional framework. In *Proceedings of EDILOG, the 6th workshop on the semantics and pragmatics of dialog*.
- Burzio, L. (1989). The morphological basis of anaphora. *Journal of Linguistics*, 27, 81–105.
- Burzio, L. (1998). Anaphora and soft constraints. In P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, & D. Pesetsky (Eds.), *Is Best Good Enough? Optimality and Competition in Syntax* (pp. 81–105). Cambridge MA: MIT Press.
- Cable, S. (2002). Hard constraints mirror soft constraints! bias, stochastic optimality and split-ergativity. ms., University of Amsterdam.
- Carden, G. & Stewart, W. (1988). Binding theory, bioprogram and creolization: evidence from hatian creole. *Journal of Pidgin and Creole Languages*, 3(1), 1–67.
- Carden, G. & Stewart, W. A. (1987). Mauritian creole reflexives: a reply to corne. *Journal of Pidgin and Creole Languages*, 4, 65–101.
- Chomsky, N. (1980). On binding. *Linguistic Inquiry*, 11, 1–46.
- Chomsky, N. (1981). *Lectures on Government and Binding Theory*. Dordrecht: Foris.
- Chomsky, N. (1982). *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge MA: MIT Press.
- Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

- Chomsky, N. (1995). *The Minimalist Program*. Cambridge MA: MIT Press.
- Clark, H. H. & Haviland, S. E. (1977). Comprehension and the given-new contrast. In R. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 11–42). Hillsdale, N.J.: Lawrence Erlbaum.
- Clark, R. & Roberts, I. (2003). A computational model of language learnability and language change. *Linguistic Inquiry*, 24, 299–345.
- Cole, P. (1981). Preface. In P. Cole (Ed.), *Radical Pragmatics*. London: Academic Press.
- Cole, P., Hermon, G., & Sung, L.-M. (1990). Principles and parameters of long-distance reflexives. *Linguistic Inquiry*, 25, 355–406.
- Corne, C. (1988). Mauritian creole reflexives. *Journal of Pidgin and Creole Languages*, 3, 69–94.
- Craig, C. (1977). *The Structure of Jacaltec*. Austin: University of Texas Press.
- Davison, A. (1999). Lexical anaphora in Hindi/Urdu. In B. Lust, K. Wali, J. Gair, & K. V. Subbarao (Eds.), *Lexical Pronouns and Anaphors in some South East Asian Languages: A Principled Typology*. Berlin: Mouton Gruyter.
- de Hoop, H. & de Swart, H. (2000). Temporal adjunct clauses in optimality theory. *Rivista di Linguistica*, 12(1), 107–127.
- Dekker, P. & van Rooy, R. (2001). Bi-directional optimality theory: An application of game theory. *Journal of Semantics*, 17, 217–242.
- Dixon, R. (1972). *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press.
- Dixon, R. (1980). *The Languages of Australia*. Cambridge: Cambridge University Press.
- Dixon, R. (1983). Nyawaygi. In R. Dixon & B. J. Blake (Eds.), *Handbook of Australian Languages*. Canberra: Australian National University Press.

- Dixon, R. (1988). *A Grammar of Boumaa Fijian*. Chicago: The University of Chicago Press.
- Eades, D. (1983). Gumbaynggir. In R. Dixon & B. J. Blake (Eds.), *Handbook of Australian Languages*. Canberra: Australian National University Press.
- Evans, G. (1980). Pronouns. *Linguistic Inquiry*, 11, 337–362.
- Everaert, M. (1991). Contextual determination of the anaphor/pronominal distinction. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 77–118). Cambridge: Cambridge University Press.
- Everaert, M. (2001). Binding theories. a comparison of ‘grammatical models’. In M. van Oostendorp & E. Anagnostopoulou (Eds.), *Progress in Grammar: Articles at the 20th Anniversary of the Comparison of Grammatical Models Group in Tilburg*. Meertens Institute Electronic Publications in Linguistics: Cambridge University Press.
- Everaert, M. & Anagnostopoulou, E. (1997). Thematic hierarchies and binding theory: evidence from Greek. In F. Corblin, D. Godard, & J.-M. Marandin (Eds.), *Empirical Issues in Formal Syntax and Semantics*. Bern: Peter Lang.
- Faltz, L. (1985). *Reflexivization: A Study in Universal Syntax*. New York: Garland.
- Farmer, A. K. & Harnish, R. M. (1987). Communicative reference with pronouns. In Verschueren & Bertuccelli-Papi (Eds.), *The Pragmatic Perspective* (pp. 547–565). Amsterdam: John Benjamins.
- Fillmore, C. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1–88). New York: Holt, Rinehart, and Winston.
- Fry, J. (2001). *Ellipsis and wa-marking in Japanese conversation*. PhD thesis, Stanford University.
- Gerdts, D. (1988). *Object and Absolutive in Halkomelem Salish*. New York: Garland.
- Giorgi, A. (1984). Towards a theory of long distance anaphors: A GB approach. *The Linguistic Review*, 4, 307–362.

- Giorgi, A. (1991). Prepositions, binding and  $\theta$ -marking. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 185–229). Cambridge: Cambridge University Press.
- Greenberg, J. (1966). *Language Universals with Special Reference to Feature Hierarchies*. The Hague: Mouton.
- Grice, H. (1957). Meaning. *Philosophical Review*, 66, 377–388.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech Acts, Syntax and Semantics 3* (pp. 41–58). London: Academic Press.
- Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Pragmatics, Syntax and Semantics 9* (pp. 113–128). London: Academic Press.
- Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge MA: Harvard University Press.
- Grimshaw, J. (1990). *Argument Structure*. Cambridge, MA: MIT Press.
- Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic Inquiry*, 28(3), 373–422.
- Grodzinsky, Y. & Reinhart, T. (1993). The innateness of binding and coreference. *Linguistic Inquiry*, 24, 69–101.
- Hara, T. (2002). *Anaphoric Dependencies in Japanese*. PhD thesis, University of Utrecht.
- Hawkins, J. A. (2002). Symmetries and asymmetries: their grammar, typology and parsing. *Theoretical Linguistics*, 28(2), 95–150.
- Hellan, L. (1988). *Anaphora in Norwegian and the Theory of Grammar*. Dordrecht: Foris.
- Hendriks, P. & de Hoop, H. (2001). Optimality theoretic semantics. *Linguistics and Philosophy*, 24(1), 1–32.
- Horn, L. (1972). *On the semantic properties of logical operators in English*. PhD thesis, University of California, Los Angeles.

- Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning Form and Use in Context: Linguistic Applications* (pp. 11–42). Washington D.C.: Georgetown University Press.
- Horn, L. R. (1989). *A Natural History of Negation*. Chicago University Press: Chicago.
- Huang, C.-T. J. (1983). A note on the binding theory. *Linguistic Inquiry*, 14, 554–561.
- Huang, C.-T. J. & Tang, C.-C. J. (1991). The local nature of the long-distance reflexives in Chinese. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 263–282). Cambridge: Cambridge University Press.
- Huang, Y. (1994). *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge: Cambridge University Press.
- Huang, Y. (2000). *Anaphora: A Cross-linguistic Study*. Oxford: Oxford University Press.
- Jackendoff, R. S. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.
- Jackendoff, R. S. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.
- Jäger, G. (2000). Some notes on the formal properties of bidirectional optimality theory. *Journal of Logic, Language and Information*, 11(4), 427–451.
- Jäger, G. (2003a). Learning constraint sub-hierarchies: The bidirectional gradual learning algorithm. In R. Blutner & H. Zeevat (Eds.), *Optimality Theory and Pragmatics* (pp. 217–242). Palgrave MacMillan.
- Jäger, G. (2003b). Maximum entropy models and stochastic optimality theory. ms. University of Potsdam.
- Jäger, G. & Rosenbach, A. (2003). The winner takes it all – almost. cumulativity in grammatical variation. ms. University of Potsdam and University of Düsseldorf.

- Jakobson, R. (1939). *Signe zéro*. In *Melanges de linguistique offerts a Charles Bally sous les auspices de la Faculte des lettres de L'Universite de Geneve par des collegues, des confreres, des disciples reconnaissants*. Geneva: Georg et cie, s.a.
- Katada, F. (1991). The lf representation of anaphors. *Linguistic Inquiry*, 22, 287–313.
- Keenan, E. L. (2000). An historical explanation of some binding theoretic facts in english. ms., [http://www.linguistics.ucla.edu /people/keen/historical.pdf](http://www.linguistics.ucla.edu/people/keen/historical.pdf), UCLA.
- Keenan, E. L. (2001). Explaining the creation of reflexive pronouns in English. ms., <http://www.linguistics.ucla.edu/people/keen/shelpaper.pdf>, UCLA.
- Kirby, S. & Hurford, J. (1997). The evolution of incremental learning: language, development and critical periods. Technical report, Language Evolution and Computation Research Unit, University of Edinburgh.
- Kiss, K. E. (1985). NP movement, operator movement, and scrambling in hungarian. In K. E. Kiss (Ed.), *Discourse Configurational Languages*, Oxford Studies In Comparative Syntax. Oxford: Oxford University Press.
- Koenig, E. & Siemund, P. (2000). *Intensifiers and reflexives: a typological perspective*, (pp. 41–74). Amsterdam: John Benjamins.
- Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1, 199–244.
- Kroch, A. & Taylor, A. (1997). Verb movement in old and middle english: dialect variation and language contact. In A. van Kemenade & N. Vincent (Eds.), *Parameters of Morphosyntactic Change*. Cambridge: Cambridge University Press.
- Lasnik, H. (1989). *Essays on Anaphora*. Dordrecht: Kluwer.
- Lebeaux, D. (1983). A distributional difference between reciprocals and reflexives. *Linguistic Inquiry*, 14, 723–730.
- Lebeaux, D. (1985). Locality and anaphoric binding. *Linguistic Review*, 4, 343–363.

- Lee, H. (2001). Quantitative variation in object marking in Korean: An experimental study.
- Levinson, S. (1987a). Minimization and conversational inference. In J. Verschueren & M. Bertuccelli-Papi (Eds.), *The Pragmatic Perspective* (pp. 61–129). Amsterdam: John Benjamins.
- Levinson, S. (1987b). Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics*, 23, 379–434.
- Levinson, S. (1991). Pragmatic reduction of the Binding Conditions revisited. *Journal of Linguistics*, 27, 107–161.
- Levinson, S. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Lidz, J. (1996). *Dimensions of Reflexivity*. PhD thesis, University of Delaware.
- Lightfoot, D. (1989). The child's trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences*, 12, 321–375.
- Magnusson, F. (1985). Skyrsla um sogulega afturbeygingu. ms., University of Iceland.
- Maling, J. (1984). Non-clause-bounded reflexives in modern Icelandic. *Linguistics and Philosophy*, 7, 211–241.
- Manzini, R. (1983). On control and control theory. *Linguistic Inquiry*, 14, 421–446.
- Manzini, R. & Wexler, K. (1987). Parameters and learnability in binding theory. In T. Roeper & E. Williams (Eds.), *Parameter Setting* (pp. 41–76). Dordrecht: Reidel.
- McCawley, J. (1978). Conversational implicature and the lexicon. In P. Cole (Ed.), *Pragmatics, Syntax and Semantics 9* (pp. 245–259). London: Academic Press.
- Mitchell, B. (1985). *Old English Syntax, Vols. I-II*. Oxford: The Clarendon Press.



- Mohanan, K. P. (1982). Grammatical relations and anaphora in Malayalam. In A. Marantz & T. Stowell (Eds.), *MIT Working Papers in Linguistics: Papers in Syntax (Volume 4)* (pp. 163–190). Cambridge, MA: MIT Press.
- Nestle, E. & Aland, B. (1983). *Novum Testamentum Graece, Nestle-Aland 26th edition*. Stuttgart: Deutsche Bibelgesellschaft.
- Niyogi, P. & Berwick, R. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy*, 20, 697–719.
- Ntelitheos, D. (2001). A constraint hierarchy approach to the different distribution of reflexives in English and Greek. MA thesis.
- O'Connor, M. (1993). Disjoint reference and pragmatic inference: Anaphora and switch reference in northern Pomo. In W. A. Foley (Ed.), *The Role of Theory in Language Description* (pp. 215–242). Berlin: Mouton de Gruyter.
- Pica, P. (1985). Subject, tense and truth: Towards a modular approach to binding. In H. G. O. J. Guéron & J.-Y. Pollock (Eds.), *Grammatical Representation* (pp. 259–291). Dordrecht: Foris.
- Pica, P. (1987). On the nature of the reflexivization cycle. In *Proceedings of NELS 17 Vol. II* (pp. 483–499). Cambridge, MA: University of Massachusetts.
- Pintzuk, S. (1991). *Phrase Structure in Competition: Variation and change in Old English word order*. Cambridge, MA: Harvard University Press.
- Platzack, C. (1987). The Scandinavian languages and the null-subject parameter. *Natural Language and Linguistic Theory*, 5, 377–401.
- Pollard, C. & Sag, I. (1992). Anaphors in English and the scope of binding theory. *Linguistic Inquiry*, 23, 261–303.
- Postal, P. (1971). *Cross-over Phenomena*. New York: Holt, Reinhart and Winston.
- Prince, A. & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science.

- Progovac, L. (1992). Relativized SUBJECT: Long-distance reflexives without movement. *Linguistic Inquiry*, 23, 671–680.
- Progovac, L. (1993). Long-distance reflexives: Movement-to-infl versus relativized SUBJECT. *Linguistic Inquiry*, 24, 755–772.
- Randriamasimanana, C. (1996). *The Causatives of Malagasy*. Honolulu: University of Hawaii Press.
- Reinhart, T. (1983). Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6, 47–88.
- Reinhart, T. (1986). Center and periphery in the grammar of anaphora. In B. Lust (Ed.), *Studies in the Acquisition of Anaphora vol. I* (pp. 123–150). Dordrecht: Reidel.
- Reinhart, T. & Reuland, E. (1991). Anaphors and logophors: an argument perspective. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 165–174). Cambridge: Cambridge University Press.
- Reinhart, T. & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry*, 24, 657–720.
- Reinhart, T. & Reuland, E. (1995). Pronouns, anaphors and case. In H. Haider, S. Olsen, & S. Vikner (Eds.), *Studies in Comparative Germanic Syntax*. Kluwer.
- Richards, N. (1997). Competition and disjoint reference. *Linguistic Inquiry*, 28, 178–187.
- Santorini, B. (1992). Variation and change in yiddish subordinate clause word order. *Natural Language and Linguistic Theory*, 10, 595–640.
- Saxena, A. (1985). Reflexivization in Hindi: a reconsideration. *International Journal of Dravidian Linguistics*, 14, 225–237.
- Schwarzschild, R. (1999). GIVENness, Avoid F and other constraints on the placement of focus. *Natural Language Semantics*, 7(2), 141–177.
- Searle, J. (1975). A taxonomy for illocutionary acts. In K. Gunderson (Ed.), *Language, Mind and Knowledge* (pp. 344–369). Minneapolis: Minnesota University Press.

- Sells, P. (1987). Aspects of logophoricity. *Linguistic Inquiry*, 18, 445–479.
- Senft, G. (1986). *Kilivila, the Language of the Trobriand Islands*. Berlin: Mouton de Gruyter.
- Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379–432, 623–656.
- Sigurðsson, H. (1990). Long-distance reflexives and moods in Icelandic. In J. Maling & A. Zaenen (Eds.), *Modern Icelandic Syntax*, Syntax and Semantics 24 (pp. 309–346). London: Academic Press.
- Smolensky, P. (1995). On the internal structure of the constraint component *Con* of *ug*.
- Stirling, L. (1993). *Switch Reference and Discourse Representation*. Cambridge: Cambridge University Press.
- Sweet, H. (1882). *An Anglo-Saxon Primer*. Oxford: Clarendon Press.
- Thráinsson, H. (1991). Long-distance reflexives and the typology of NPs. In J. Koster & E. Reuland (Eds.), *Long-distance Anaphora* (pp. 49–79). Cambridge: Cambridge University Press.
- Tryon, D. (1970). *Conversational Tahitian*. Canberra: Australian National University Press.
- van der Does, J. & de Hoop, H. (1998). Type-shifting and scrambled definites. *Journal of Semantics*, 15(4), 393–416.
- Visser, F. (1963). *An Historical Syntax of the English Language*. Leiden: Brill.
- Wali, K. (1989). *Marathi Syntax: A Study of Reflexives*. Patiala, India: dian Institute of Language Studies.
- Wang, J. & Stilings, J. (1984). Chinese reflexives. In X. Li (Ed.), *Proceedings of the 1st Harbin Conference on Generative Grammar* (pp. 100–109). Harbin, China: Heilongjiang University Press.
- Wilkins, W. (1988). Thematic structure and reflexivization. In W. Wilkins (Ed.), *Thematic Relations*, Syntax and Semantics 21 (pp. 191–213). London: Academic Press.

- Williams, E. (1977). Discourse and logical form. *Linguistic Inquiry*, 8, 101–139.
- Wilson, C. (2001). Bidirectional optimization and the theory of anaphora. In G. Legendre & S. Vikner (Eds.), *Optimality-theoretic Syntax* (pp. 465–507). Cambridge: MIT Press.
- Yang, C. D. (2000). *Knowledge and Learning in Natural Language*. PhD thesis, MIT.
- Zeevat, H. (2001). The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics*, 17, 243–262.
- Zeevat, H. (2002). Double bias. In Alberti, Balogh, & Dekker (Eds.), *Proceedings of the Seventh Symposium on Logic and Language, Pecs 2002* (pp. 173–181). Pecs.
- Zeevat, H. & Jäger, G. (2002). A statistical reinterpretation of harmonic alignment. In D. de Jongh, M. Nilsenova, & H. Zeevat (Eds.), *Proceedings of the 4th Tblisi Symposium on Logic, Language and Linguistics. Amsterdam/Tblisi 2002* (pp. 173–181). Tblisi.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison Wesley.

## General Index

- A-binding, 10  
A-chains, 29  
A-positions, 10, 17, 29, 51, 61, 71, 72, 104  
Antisynonymy, 71–73, 104, 134, 143, 150  
asymmetric bidirectional optimality, 90  
bias, 7, 77  
bias constraints, 103, 109, 110, 112–114, 116–119, 121, 133, 135, 137, 139–141, 144, 146, 148–150, 152–156, 159, 160, 162–164  
Bidirectional Gradual Learning Algorithm (BiGLA), 6, 90–93, 95–98, 101, 103, 104, 106, 107, 110, 114, 116, 118, 121, 124, 129, 136, 145, 153  
bidirectional learning, 7, 8, 91, 96, 98, 102, 103, 109, 110, 119, 131–135, 137, 138, 142, 144, 148, 152, 154, 156, 158, 159, 162–164  
bidirectional Optimality Theory, 6, 8, 56–58, 61, 68, 70, 78  
bidirectional optimization, 52, 54, 55, 60, 69, 76, 78, 85, 87, 88, 90, 97, 113, 120, 125–127, 132–134, 136–138, 150, 151, 159, 163, 164  
Binding Principles, 11, 12, 21, 28, 32, 35, 43  
Principle A, 10, 11, 16, 18, 20, 21, 24, 28–30, 32, 44, 46, 130, 151  
Principle B, 10, 11, 16–19, 24, 28–30, 32  
Principle C, 10, 11, 13, 14, 16, 18, 19, 26, 32, 33  
Principles C1 & C2 (Lasnik), 14  
blocking, 42, 55, 56, 78–80, 87, 88, 90, 136, 150  
c-command, 29, 151  
conversational implicature, 7, 8, 33, 36–45, 47–51, 54, 62, 67, 70, 110, 143, 151  
Cooperative Principle (Grice), 35–37  
differential case marking (DCM), 74, 78, 79, 81, 85, 89, 90, 92, 94–96  
Disjoint Reference Presumption (DRP), 45, 163, 165  
division of pragmatic labor (Horn), 38, 42, 54, 69, 73, 165  
evolutionary OT, 93, 98, 103, 142  
ganging-up cumulativity, 113, 114, 118, 119, 133, 135, 141, 164  
General Discourse Principle (Chomsky), 15, 16, 26, 33  
Government & Binding Theory, 9  
Gradual Learning Algorithm (GLA), 6, 8, 69, 81, 84, 85, 89, 90, 118

grammaticalization, 7, 8, 35, 43,  
 44, 47, 50–52, 59, 63, 64,  
 68–70, 73, 81, 84, 89, 95,  
 102, 103, 129, 138, 150, 157–  
 160, 162, 165

harmonic alignment, 74, 75  
 Horn-scale, 40, 41, 67

I-principle, 38–40, 42, 44, 45  
 Iterated Learning Model (ILM), 87,  
 88

logophoricity, 26, 48–50, 66, 158–  
 160, 163

long-distance Anaphors (LDAs), 20–  
 26, 31, 49, 73, 133, 151, 154,  
 155, 157–160

M-principle, 38, 42, 46–48, 56, 58,  
 59, 68, 72, 73, 90, 103, 134,  
 143

Minimalism, 33, 53

Pattern Generalization, 71–73, 104,  
 133, 134, 136, 138–141, 143

predicate meaning/reflexivizing strat-  
 egy correlation, 143, 151

Principles & Parameters Theory, 9

Q-principle, 38–42, 48, 58, 60

R-principle, 38

radical pragmatics, 37, 56

relative harmony, 53

SE anaphora, 25–28, 30, 31, 50, 60,  
 61, 66, 67, 104, 133, 138,  
 142, 144–146, 149, 155, 158–  
 160

stochastic OT, 6, 8, 69, 81, 82, 84,  
 161, 163

Thematic Heirarchy Condition, 19,  
 133, 151, 163

weak bidirectional optimality, 57,  
 59, 68–70, 72, 73, 103, 134,  
 143

## Index of Authors

- Aikawa, T., 31  
Aissen, J., 74–76, 78, 85, 90, 92, 98–100  
Anagnostopoulou, E., 19, 20, 154  
Atlas, J.D., 6, 39, 42, 56, 133, 163  
Austin, P., 64, 74  
Bar-Hillel, Y., 48  
Battistella, E., 21  
Beaver, D., 90  
Benedicto, E., 23  
Berwick, R., 71  
Blutner, R., 6, 8, 52, 54, 55, 57–59, 68, 69, 72, 73, 76, 90, 103, 129, 133, 134, 143, 163  
Boersma, P., 6, 8, 52, 69, 81, 82, 84  
Bouchard, D., 10, 21, 27, 157  
Bresnan, J., 54, 59  
Briscoe, T., 71  
Buchwald, A. et al., 6, 59, 60, 90, 105  
Burzio, L., 10, 22, 25, 26, 54, 60, 159–161  
Cable, S., 7, 8, 52, 69, 73, 81, 85–87, 89, 90, 92, 93, 103, 126, 162  
Carden, G., 17, 18, 45, 47, 63, 96, 103  
Carnap, R., 48  
Chomsky, N., 6, 8, 9, 12, 15, 16, 26, 27, 32–34, 43, 44, 49, 53, 128, 134, 160, 161, 163  
Clark, R., 39, 71  
Cole, P., 22, 31, 37  
Corne, C., 63  
Craig, C., 77  
Dahl, Ö., 76  
Davison, A., 157, 158  
de Hoop, H., 54  
de Swart, H., 54  
Dekker, P., 56  
Dixon, R., 17, 64, 65, 74  
Eades, D., 64  
Evans, G., 15  
Everaert, M., 9, 19, 20, 30, 154  
Faltz, L., 17, 18, 22, 46, 61, 66, 67, 105  
Farmer, A.K., 29, 44, 45, 110  
Fillmore, C., 19  
Fry, J., 77  
Gerdts, D., 77  
Giorgi, A., 19  
Greenberg, J., 76  
Grice, H.P., 35–37, 41, 54, 56, 72  
Grimshaw, J., 19, 20, 54, 151  
Grodzinsky, Y., 16  
Hara, T., 23  
Harnish, R.M., 29, 44, 45, 110  
Haviland, S.E., 39  
Hawkins, J.A., 162  
Hayes, B., 82  
Hellan, L., 157  
Hendriks, P., 54  
Hermon, G., 22, 31

- Horn, L., 6, 37, 38, 40, 42, 52, 54,  
56, 69, 73, 98, 103, 133, 163,  
165
- Huang, C.-T.J., 21, 22
- Huang, Y., 6, 11, 15, 19, 21–23, 26,  
31, 34, 70, 110, 142, 143,  
155, 156, 160, 161
- Hurford, J., 88
- Jäger, G., 6–8, 52, 54, 57, 69, 73,  
74, 76, 77, 79–81, 87, 89–  
93, 96, 98, 100, 103, 113,  
121, 125–127, 133, 134, 138,  
141, 162
- Jackendoff, R., 19, 151
- Jakobson, R., 76
- Kaplan, R., 59
- Katada, F., 22
- Keenan, E., 17, 61, 70–73, 89, 104,  
133, 134, 136, 143
- Kirby, S., 88
- Kiss, K.E., 19
- Koenig, E., 142, 151
- Kroch, A., 130
- Lasnik, H., 10, 13–15, 161, 162
- Lebeaux, D., 21, 157
- Lee, H., 77, 90
- Levinson, S., 6–8, 17, 18, 26, 29,  
32, 34, 35, 37–52, 54, 56,  
58–61, 63–65, 67–74, 89, 90,  
96, 98, 103–105, 110, 125,  
133, 134, 143, 151, 158, 159,  
161–163
- Lidz, J., 31, 142
- Lightfoot, D., 71
- Magnusson, F., 159
- Maling, J., 159
- Manzini, R., 10, 21, 23, 24
- McCawley, J., 42, 57
- Mitchell, B., 17, 46, 61
- Mohanan, K.P., 21
- Niyogi, P., 71
- Ntelitheos, D., 151
- O'Connor, M., 26, 48
- Pica, P., 21, 22
- Pintzuk, S., 130
- Platzack, C., 71
- Pollard, C., 21
- Postal, P., 19
- Prince, A., 6, 25, 52, 53, 72, 75
- Progovac, L., 21, 23
- Randriamasimanana, C., 19
- Reinhart, T., 6, 8, 9, 16, 22, 26–  
29, 31, 32, 53, 67, 134, 141,  
142, 151, 163
- Reuland, E., 6, 8, 9, 16, 22, 26–  
29, 31, 32, 53, 67, 134, 141,  
142, 151, 163
- Richards, N., 161
- Roberts, I., 71
- Rosenbach, A., 113, 141
- Sag, I., 21
- Santorini, B., 130
- Saxena, A., 157
- Schwarzschild, R., 66
- Searle, J., 40
- Sells, P., 21
- Senft, G., 64
- Shannon, C., 6, 74, 133
- Siemund, P., 142, 151



Sigurðsson, H., 21, 25, 26, 160  
Smolensky, P., 6, 25, 52, 53, 72, 75,  
92  
Stewart, W.A., 17, 18, 45, 47, 63,  
96, 103  
Stillings, J., 21  
Stirling, L., 26, 48  
Sung, L.-M., 22, 31  
Sweet, H., 46  
  
Tang, C.-C.J., 22  
Taylor, A., 130  
Thráinsson, H., 10, 21, 26, 158  
Tryon, D., 17  
  
van der Does, J., 54  
van Rooy, R., 56  
Visser, F., 17, 18, 46, 47, 61  
  
Wali, K., 22  
Wang, J., 21  
Wilkins, W., 19  
Williams, E., 157  
Wilson, C., 6, 60, 90, 105  
  
Xu, Y., 21  
  
Yang, C.D., 71  
  
Zeevat, H., 7, 8, 52, 54, 69, 73, 74,  
76, 77, 79, 80, 89, 90, 92,  
93, 101, 103, 105, 121, 126,  
127, 162  
Zipf, G.K., 37

## Index of Languages

- Bislama, 47, 63  
Chinese, 15, 20–22, 24, 26, 155, 161  
Dhargari, 74  
Dutch, 28, 30, 51, 73, 104, 133, 134, 138, 141–144  
Dyirbal, 74  
Efik, 67  
English, 34  
    Middle English, 18, 24, 47, 51, 61, 137  
    Modern English, 13, 15–18, 27, 30  
    Old English, 17, 34, 44–47, 50, 51, 61–63, 66, 82, 103, 104  
Ewe, 67  
Fijian, 17, 45, 46, 65  
French, 27  
Frisian, 30  
German, 18, 31, 90, 142  
Greek  
    Classical Greek, 13, 15  
    Modern Greek, 19, 20, 31, 133, 151–154  
Guadeloupe, 17, 45  
Gumbaynggir, 64  
Guugu Yimithirr, 17, 34, 45, 46, 64  
Haitian Creole, 18, 24, 44–46, 63, 64, 66, 165  
Halkomelem, 77  
Hindi/Urdu, 158  
Hungarian, 133  
Icelandic, 20, 21, 24–27, 49, 66, 142, 158–160  
    Old Icelandic, 26, 159, 160  
Igbo, 66  
Italian, 27  
Jacaltepec, 77  
Japanese, 23, 26, 31, 151, 155  
Jiwarli, 64  
Kilivilla, 64  
KiNubi, 17, 45  
Korean, 26, 31, 155  
Kriyol, 63  
Lakhota, 66  
Latin, 23  
Malagasy, 19, 31, 133, 151  
Marathi, 22  
Martinique Creole, 47, 63  
Mauritian Creole, 47, 63  
Negerhollands, 63  
Norwegian, 157  
Nyawaygi, 64  
Padovano, 31, 151  
Palenquero, 17, 45  
Proto-Indo-European, 66  
Spanish, 27, 31, 151  
Swedish, 77  
Tahitian, 17, 64  
Thai, 13–15, 82  
Vietnamese, 13, 14

Yiddish, 75  
Yoruba, 66