A note on acquisition in frequency-based accounts of Binding Phenomena

Jason Mattausch and Insa Gülzow

This article addresses the so-called 'pronoun interpretation problem' or 'delay of Principle B effect' – an observation in the study of language acquisition that challenges classical Binding Theory. We show that a recent, frequentist theory of binding which is empirically superior to the classical Binding Theory can, with a minor adjustment, address the pronoun interpretation problem and thus explain why children acquire interpretational restrictions on pronouns later than they acquire such restrictions on reflexives and why the acquisition of interpretational restrictions lags behind restrictions on generation.

1. Introduction

The purpose of this article is to reconcile a recent, novel approach to binding phenomena with data from acquisition experiments, namely the so-called 'pronoun interpretation problem' or 'delay of Principle B effect', which has been noted in several acquisition studies, e.g. Wexler & Chien (1990) and Grimshaw & Rosen (1990).

The structure of the paper is as follows. In the following section, classical binding theory is introduced and we discuss three problems related to the theory. We then outline a recent alternative to classical binding theory, which solves two of these three problems. Section 4 discusses a solution to the third problem.

2. Three challenges to classical binding theory

The following examples illustrate a very common pattern in natural language binding phenomena.

(1) a. **He*$_i$ *pleases him*$_i$
b. *He*$_i$ *pleases him*$_j$

c. *He$_i$ pleases himself$_i$*
d. **He$_i$ pleases himself$_j$*

Below are two principles of Chomsky's Binding Theory (BT), the most well-known approach to explaining the pattern manifested in (1), (Chomsky 1981).

(2) a. Binding Principle A: a reflexive must be locally bound.
b. Binding Principle B: a pronoun must be locally free.

Principles A and B account for the ungrammaticality of (1a), where a pronoun is locally bound and for the ungrammaticality of (1d), where a reflexive is locally free.[1]

The BT analysis of the pattern exemplified in (1) both undergenerates and overgenerates. Firstly, as pointed out in Levinson (2000), there are a considerable number of languages which appear to lack morphological means of encoding reflexivity altogether and use pronouns reflexively, thus disobeying Principle B systematically and obeying Principle A only vacuously.

One example is English itself, though not its modern form. Specifically, evidence from Old English (cf. Visser 1963: 420–439; Mitchell 1985: 115–189; Keenan 2000, 2001) shows that the opposition between the OE pronoun *hine* and the emphatic *hine selfne* is not comparable to the opposition between the modern counterparts *him* and *himself*, since *hine* could appear locally bound and *hine selfne*, though often used as a reflexive, did not necessarily take a local antecedent.

(3) Old English (Siemund 2000)
Hine$_i$ he$_i$ bewerað mid wæpnum.
Him he defended with weapons
'He defended himself with weapons.'

(4) Old English (Mitchell 1985: 115)
Moyses$_i$, se ðe wæs God$_j$ sua weorð ðæt he$_i$ oft wið
Moses he who was to-God so dear that he often with
hine selfne$_j$ spræc.
him self spoke
'Moses was so dear to God that he often spoke with him.'

A second challenge to BT can be illustrated by imagining a hypothetical language which manifested the pattern shown below in (5).

(5) 'Anti-English'

a. *He$_i$ pleases him$_i$*
b. **He$_i$ pleases him$_j$*
c. **He$_i$ pleases himself$_i$*
d. *He$_i$ pleases himself$_j$*

We call this hypothetical language 'Anti-English' because it exhibits exactly the opposite pattern as modern English in that the *self* -marked form is only grammatical with a non-local antecedent whereas the bare pronoun demands a local one. One working in the BT framework would have no problem accounting for such a pattern – he would surely just say that, in Anti-English, *him* is a reflexive whereas Anti-English *himself* is a pronoun. The question then becomes why there are no languages like Anti-English, i.e., why do languages mark pronominal objects of reflexive predicates instead of marking those of non-reflexive ones? Standard BT offers no answer to this question.[2]

Finally, classical BT faces difficulty accounting for data observed in the study of language acquisition. Several studies have observed a 'pronoun interpretation problem' or 'delay of Principle B effect', whereby children (a) appear to interpret and produce reflexives in accordance with the binding principles, (b) appear to produce pronouns in accordance with the binding principles and (c) do not appear to interpret pronouns in accordance with the binding principles, but rather interpret pronouns reflexively about 50% of the time in experiments (e.g. Shipley & Shipley 1969; Charney 1980; Chiat 1981; Loveland 1984; Chien & Wexler 1990; Grimshaw & Rosen 1990; Girouard, Richard & Décarie 1997). The issue is discussed in more detail below, but it shall suffice here to note that if, as classical BT holds, Principles A and B are innate, universal tenets of grammar then the 'delay of Principle B' effect is not predicted to occur.

We shall proceed by outlining an alternative to classical BT due to Mattausch (2004, 2006), which addresses the problems of undergeneration and overgeneration, then turn to a novel solution to the acquisition problem based on that approach.

3. An alternative to BT

This section outlines the approach to binding phenomena proposed by Mattausch (2004, 2006) and shows how the approach addresses the first two challenges to classical BT discussed above. The account is based on Bidirectional-Evolutionary Optimality Theory, which is introduced below.

3.1. Optimality Theory

Optimality Theory (OT) is a theory of grammar that gives up the idea of absolute principles of grammar in favor of conflicting, violable constraints, which can be ranked in various possible ways to reflect their strength in a particular language. In OT, a certain input gets associated with a multitude of possible outputs or candidates. Each candidate is then evaluated with respect to a series of ranked constraints, of which there are two basic types – *faithfulness constraints*, which penalize divergence of the output candidate from the original input and *markedness constraints*, which militate against certain features or properties of the output. 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.

3.2. Bidirectional Optimality Theory

In generative grammars whose essence is to produce morphological or syntactic expressions for some underlying meaning, the definition of optimality is as follows.

(6) Optimality (production)

A form *f* is an optimal expression, given a meaning *m*, iff there is no *f'* such that *f'* is more harmonic than *f* (write: *f'* > *f*), given *m* as an input.

In comprehension grammars whose essence is to interpret morphological or syntactic expressions, the definition of optimality is as below.

(7) Optimality (comprehension)

A meaning *m* is an optimal interpretation, given a form *f*, iff there is no *m'* such that *m'* > *m*, given *f* as an input.

Bidirectional OT (championed by Blutner 2000; Wilson 2001; Zeevat 2001; Jäger 2003a) is a variation of OT meant to incorporate both production and comprehension aspects of language into one grammar and capture the interdependency of the two. The issue of interdependency is crucial, since it is commonsensical to capture the idea that, in a communication situation, an expression should, first and foremost, allow the hearer to recover the intended meaning of the expression. Such an idea is captured by formulating a definition of bidirectional optimality as below.

(8) Bidirectional optimality (Jäger, 2003a: 19)

a. A form-meaning pair ⟨*f,m*⟩ is hearer optimal iff there is no pair ⟨*f,m'*⟩ such that ⟨*f,m'*⟩ > ⟨*f,m*⟩.

b. A form-meaning pair ⟨*f,m*⟩ is optimal iff either
   (i) ⟨*f,m*⟩ is hearer optimal and there is no distinct pair ⟨*f',m*⟩ such that ⟨*f',m*⟩ > ⟨*f,m*⟩ and ⟨*f',m*⟩ is hearer optimal, or
   (ii) no pair is hearer optimal and there is no distinct pair ⟨*f',m*⟩ such that ⟨*f',m*⟩ > ⟨*f,m*⟩.

Note that the definition above contains a recoverability restriction for generative optimality: forms are disqualified as candidates when they are not optimally recoverable as the intended meaning and at least one other form is. Where a form is disqualified due to the recoverability restriction, it is said to be *blocked*.

3.3. Addressing undergeneration: stochastic, bidirectional learning

One key to addressing the problem of undergeneration is a 'stochasticization' of OT and a learning theory that goes along with it. Both are introduced below, followed by an illustration of how they are useful in formulating a more descriptively adequate account of binding phenomena.

3.3.1. Stochastic Optimality Theory

The *Stochastic OT* (StOT) of Boersma (1998) and Boersma & Hayes (2001) is a variation of standard OT in which a 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, one 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.

Constraint rankings in StOT are continuous, each constraint 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.

StOT also 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 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 cumulative distribution function of a normal distribution[3] such that the mean $\mu=0$ and the standard deviation $\sigma=2\sqrt{2}$, as is roughly depicted in Figure 1.

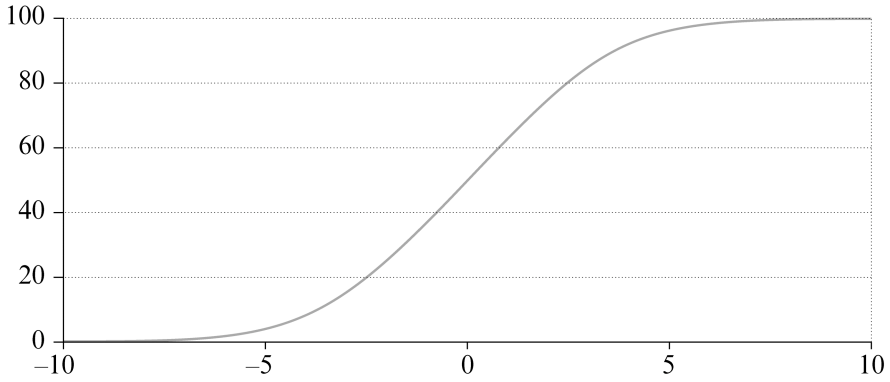

Figure 1. P($C_1 \gg C_2$), per $C_1 - C_2$ (in %)

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 or more. On the other hand, true free variation is predicted where two constraints have exactly the same ranking value. 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. If C_1 is higher ranked than C_2, there is a preference for the C_1-favored candidates. If the difference in ranking values is 2, the chance that C_1 will outrank C_2 for any given evaluation is about 76%. A difference of 5 units corresponds to a 96% chance that C_1 will outrank C_2, and so on.

3.3.2. Bidirectional learning

Boersma's Gradual Learning Algorithm (GLA) is a method of systematically generating a StOT 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 StOT 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.[4] A grammar can be said to have converged just in case further observations no longer induce significant adjustments of the learner's hypothetical grammar.

Jäger (2003a) proposes a bidirectional version of the GLA, called the Bidirectional Gradual Learning Algorithm, or BiGLA. 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, ‹*f,m*›, 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*.[5] 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.

(9) BiGLA (Jäger, 2003a: 20–21)

a. Initial state
   All constraint values are set to 0.

b. Step 1: Observation
   The algorithm is presented with a learning datum, a fully specified input-output pair ‹*f,m*›.

c. 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 $C_1 \gg \ldots \gg C_n$. 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*.

d. Step 3: Comparison
   If *f'* = *f*, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum ‹*f,m*› with the hypothetical pair ‹*f',m*›.
   If *m'* = *m*, nothing happens. Otherwise, the algorithm compares the constraint violations of the learning datum ‹*f,m*› with the hypothetical pair ‹*f,m'*› .

e. Step 4: Adjustment
   All constraints that favor ‹*f,m*› over ‹*f',m*› are increased by the plasticity value. All constraints that favor ‹*f',m*› are decreased by the plasticity value.
   All constraints that favor ‹*f,m*› over ‹*f,m'*› are increased by the plasticity value. All constraints that favor ‹*f,m'*› are decreased by the plasticity value.

f. Final state
   Steps 1–4 are repeated until the constraint values stabilize.

Jäger's idea of bidirectional learning is a crucial step in formulating a truly frequentist theory of grammar, since it allows a learner to possess a sensitivity to statistical states of affairs in the data from which he is learning, namely the relative frequency of messages that speakers convey and the relative frequency of the signals they use to convey them. Below in section 3.3.3 we sketch a solution to the undergeneration problem in classical BT based on bidirectional, stochastic OT and section 3.4 shows how such a frequentist account can also address the overgeneration problem. Finally, section 4 discusses the pronoun interpretation problem and shows how it too can be solved.

3.3.3. Addressing the undergeneration problem

The undergeneration problem that faces classical BT apparently stems from the fact that while there is obviously some force at work that militates against pronouns appearing locally bound, and some force against *self*-marked pronouns appearing locally free, these forces are not absolute. The only means of capturing 'non-absoluteness' of any constraint in OT is to postulate some conflicting constraint. For now, we shall simply imagine that a grammar consists of two constraints which mimic the force of Principles

A and B, as well as two conflicting constraints which diametrically oppose those forces.

(10) a. **self,dis*: *self*-marked pronouns are not locally disjoint.
b. **self,co: self*-marked pronouns are not locally conjoint.
c. **pro,co*: bare pronouns are not locally conjoint.
d. **pro,dis*: bare pronouns are not locally disjoint.

The constraints above are quite non-standard, for they are obviously neither markedness constraints nor faithfulness constraints. Mattausch (2004, 2006) however, advocates their invocation and proposes the moniker *bias constraints* – constraints that refer to each possible form-meaning pair and, as a set, simply behave like an OT 'counting machine' when coupled with GLA style learning in the sense that they will simply reflect statistical states of affairs in a training corpus by their relative rank to one another.

One should be able to see that any categorical pattern of binding behavior can be captured by some ranking of the bias constraints above.

Table 1. A partial factorial typology, per (10)

Constraint ranking	Language type
*self,dis, *pro,co ≫ *self,co, *pro,dis	Modern English
*self,dis, *self,co ≫ *pro,co, *pro,dis	Old English (no reflexives)
*pro,dis, *self,co ≫ *pro,co, *self,dis	Anti-English
*pro,dis, *pro,co ≫ *self,co, *self,dis	Anti-Old English (no simplex pronouns)

Moreover, we can employ stochastic OT to illustrate how languages like Middle English, where pronouns and reflexives were both attested but did not appear in complementary distribution, can be represented by a stochastic ranking of the constraints under consideration. Consider an extreme example where a language made no distinction at all between pronouns and reflexives with respect to where they could appear, and no distinction in their interpretation. We will take it for granted that sentences in which the subject and object refer to distinct entities constitute the vast majority – we'll say 98% – of sentences used by speakers of all grammars. (Note that this assumption is similar to the Disjoint Reference Presumption (DRP) of Farmer and Harnish (1987) but rather than seeing it as a interpretational presumption made by language users, we take it as a simple fact of life

about language use.) A speaker who spoke the hypothetical language we are considering would produce corpus frequencies like those below.

Table 2. Hypothetical frequencies of pronoun/reflexive distribution

	pro	pro+self	% marked
disjoint	49 %	49 %	50 %
conjoint	1 %	1 %	50 %

We can use the frequencies in Table 2 to simulate a grammar learned based on those frequencies. Feeding BiGLA with twenty thousand form-meaning pairs drawn at random based on the frequencies in Table 2 resulted in the learning curves in Figure 2.[6]

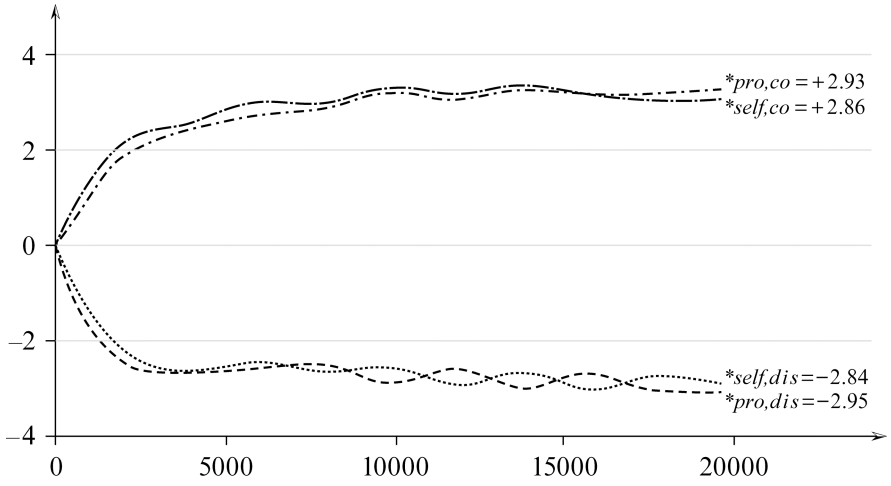

Figure 2. Learning curves (20K inputs) per Table 2

The resulting grammar shown in Figure 2 is roughly what one should expect under the circumstances. The large gap between the highly ranked constraints **pro,co* and **self,co* on the one hand and the low ranked **pro,dis* and **self,dis* on the other represent the preference for interpreting the arguments of predicates as disjoint. Note that this is basically a stochastic version of the Disjoint Reference Presumption, 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 a statistical asymmetry in the training corpus and the application of hearer-mode learning to constraints which 'record' that asymmetry.

On the other hand, the constraints **pro,co* and **self,co* have been learned as having almost exactly the same ranking value, and the same is true for **pro,dis* and **self,dis*. This reflects the fact that, in the training corpus, reflexives and pronouns were in perfectly parallel distribution.

Various degrees of variation can be also be captured, though we leave the reader to prove to him- or herself that the four constraints at our disposal in a StOT framework give us all we need to easily handle the problem of undergeneration that faced standard BT.[7]

3.4. Addressing overgeneration: Evolutionary Optimality Theory

With respect to the overgeneration problem in standard BT, one faces the task of explaining why there are no languages like 'Anti-English', where morphologically complex expressions play the role of pronouns and morphologically simplex expressions are reflexive. Fortunately, aside from the advantages already mentioned, stochastic, bidirectional OT and bidirectional learning offer an interesting opportunity to describe language change. Moreover, explanations about the direction of language change can be found when one considers what types of constraints grammars consist of and how these constraints interact.

The Iterated Learning Model (ILM) of language evolution due to Kirby & Hurford (1997) 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. In the context of bidirectional gradual learning of a StOT grammar, the first-generation learner would be exposed to a set of corpus frequencies, he would adjust his grammar accordingly until it converged into an appropriate set of ranking values. He would produce his own speech in accordance with the grammar he had acquired and the frequencies of his own speech would serve as the corpus frequencies for the second-generation learner. Thus, per the ILM, a learner who acquired the grammar in Figure 2 would himself become a 'teacher' to the next generation of learners

and the frequencies that he produced would serve as a training corpus for others. The actual simulated output frequencies for a speaker whose grammar was the one in Figure 2 are given in Table 3.

Table 3. Output frequencies per Figure 2

	pro	pro+self	% marked
disjoint	50.95 %	47.05 %	48 %
conjoint	0.88 %	1.12 %	56 %

There has obviously been some cross-generational fluctuation between the (hypothetical) grammar that generated the training corpus and the learned grammar. This is not at all uncommon; in fact, perfect statistical replication of a non-categorical marking pattern from one generation to the next is very rare. Based on the frequencies above, we see that the first generation learning has taken a step toward the Modern English pattern, since *self*-marked outputs have decreased for disjoint inputs and increased for conjoint ones. However simulations of language evolution from a neutral starting point of grammars comprised of the bias constraints above were unpredictable. There are three possible scenarios: (a) evolution into English, (b) evolution into Anti-English and (c) neither (a) nor (b), i.e., persistent variation. Conducting multiple simulations showed that all of these results were achievable and thus, so far, nothing explains why Anti-English-type grammars are unattested in natural language.

However, adding a markedness constraint to represent some universal force of structural economy causes this picture to change significantly.[8] Let us assume that an additional constraint represents a universal force of articulatory economy.

(11) **Struct*: Avoid morphological structure.

The inclusion of a constraint like **Struct* will be very significant. 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. With respect to the case at hand, (ignoring blocking effects 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* outranks **self,co*, but rather to the probability that **pro,co* outranks both **self,co* and **Struct*. Moreover, because of the

mechanics of the (Bi)GLA, there will be a strict relationship between the ranking value of **Struct* and the various bias constraints.[9] 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. (Jäger & Rosenbach 2003 call this effect 'ganging-up cumulativity' – each constraint is relevant to the evaluation regardless of its ranking value.[10])

To see the difference between learning a grammar with bias constraints only, as above, and a grammar with bias constraints and a markedness constraint, we can again feed BiGLA with twenty thousand form-meaning pairs drawn at random based on the frequencies in Table 2. The result was the learning curves in Figure 3.

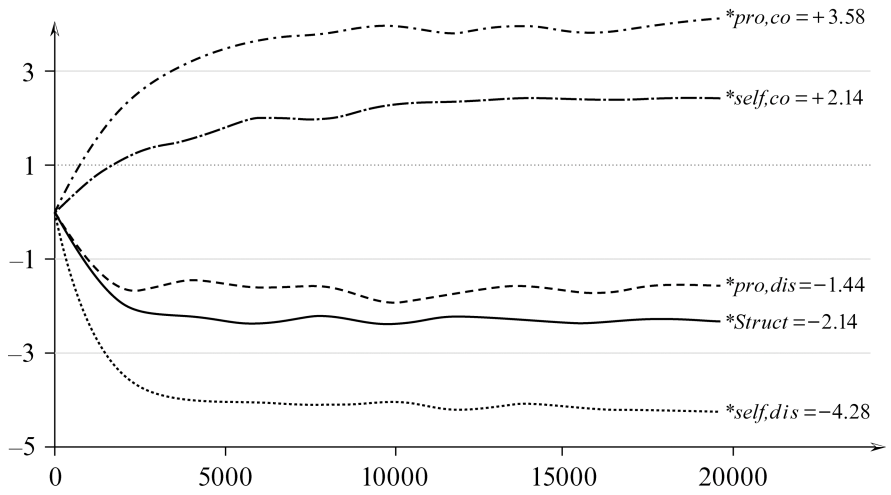

Figure 3. Learning curves (20K inputs) per Table 2

While it may be diffcult to see with the naked eye, the learned grammar in Figure 3 is very different from the one in Figure 2. Briefly stated: because **Struct* is strictly a generative constraint (i.e., it is neither promoted nor demoted in the hearer-mode), hearer-mode and speaker-mode will be learning a different number of constraints. Hearer-mode learning will be struggling to keep ranking values exactly as they were in Figure 2 whereas speaker-mode learning will be struggling to find a proper balance between the bias constraints and the markedness constraint. But a proper balance cannot be found and the compromise that is reached will favor generative accuracy for the more common type of learning data, i.e., form-meaning

pairs where the subject and object are disjoint.[11] The resulting output frequencies are in Table 4.

Table 4. Output frequencies per Figure 3

	unmarked	marked	% marked
disjoint	52.41 %	45.59 %	46.5 %
conjoint	.64 %	1.36 %	68 %

One can see that – in the spirit of Shannon's (1948) 'optimal coding' and what Horn (1984) called a 'division of pragmatic labor' – marked forms have gravitated toward rare meanings. The marked-forms-for-rare-meanings pattern taking shape here can be seen as a direct consequence of four things: (a) bias constraints (b) markedness constraints (c) the mechanics of the GLA and (d) the bidirectional application of those mechanics.

The new asymmetry that has shown up in the first-generation learner's corpus frequencies will 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. Without going into detail, the inevitable result of evolutionary simulations using a grammar with bias constraints in (10), plus ***Struct***, beginning with the corpus frequencies in Table 2 is illustrated in Figure 4.[12]

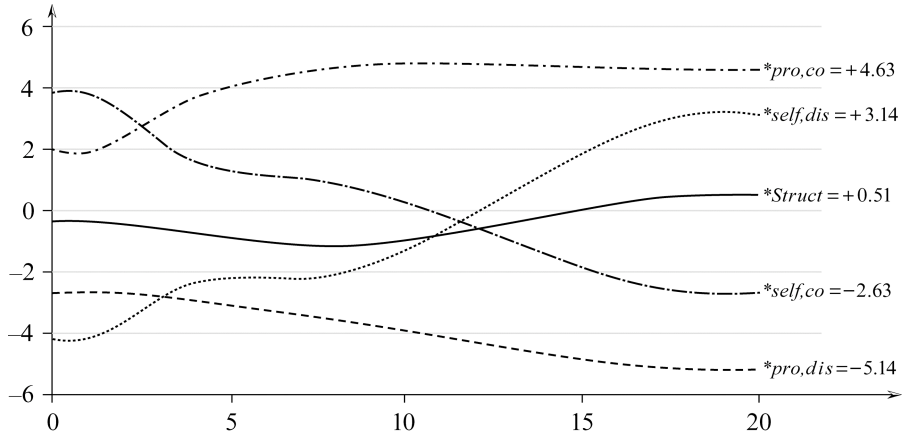

Figure 4. Evolution (20 generations)

The evolved grammar strictly follows the Principle A and B patterns of Modern Standard English, see Table 5.

Table 5. Output frequencies (100th generation)

	pro	pro+*self*	% marked
disjoint	98 %	0 %	0 %
conjoint	0 %	2 %	100 %

This result was, as noted, the only result achievable using the constraints and frequencies under consideration and thus the overgeneration problem of standard BT can be solved by considering a frequentist, evolutionary account of binding phenomena that hinges on the interaction of bias constraints and markedness constraints, which guarantees a marked-form-for-rare-meaning strategy.

4. The pronoun interpretation problem

As already mentioned above, it has been demonstrated in studies conducted within the standard BT framework that children acquiring English as their first language disobey Binding Principle B for a relatively long time in their interpretation of pronouns. Chien & Wexler (1990) have shown that English children younger than four years of age have a great tendency to interpret sentences like (12a) as if they meant (12b).

(12) a. *Mama Bear is touching her*.
b. *Mama Bear is touching herself.*

In an experiment testing English children's knowledge of the Binding Principles, children were shown pictures with the characters Goldilocks and Mama Bear. If shown a picture in which Mama Bear is touching herself, children younger than four years of age tend to answer the question *Is Mama Bear touching her?* with *yes*. Performance becomes better with increasing age, although children between five and six years of age still perform at chance level and children in the age group between six and seven years of age reacted in a target like manner only in 76% of the cases (Chien & Wexler 1990: 269, 273).[13] Similar results are reported by Grodzinsky & Reinhart (1993) who tested ungrammatical coreference in sentences like

Grover touches him with pictures in which Grover is touching someone else. No evidence exists suggesting that children exhibit the same disobedience of Principle B in their production of anaphoric expressions. Obviously, the results of such comprehension studies present a challenge to classical BT. Some previous attempts to resolve the problem are outlined below.

Chien & Wexler (1990) for instance claim that until a comparatively late age, English children overgeneralize or misinterpret the rare occurrences of a coreferential interpretation of a personal pronoun and a preceding noun phrase. They follow Reinhart (1983, 1986) in arguing that children know Principle B but lack a Pragmatic Principle P. In (13a) both *he* and *him* are taken to be John, thus *he* and *him* are coreferential.[14] The indexing in these sentences must be as represented in (13b) and (13c), as (13d) would suggest that *him* is referentially dependent on *he* which would violate Principle B.

(13) a. *That must be John*.
b. *That$_i$ must be John$_j$*.
c. *At least he$_i$ looks like him$_j$*.
d. **At least he$_i$ looks like him$_i$*.

Before children have a Pragmatic Principle P they are unable to realize that coreference of *he* and *him* in (13c) is only possible in very specific contexts. Therefore, they overgeneralize this rather rare occurrence and also allow non-target coreference in other contexts.

Grodzinsky & Reinhart (1993) offer a solution that is based on the subtle interpretational differences of a sentence such as (14a), represented in (14b) and (14c).[15]

(14) a. *Alfred thinks he is a great cook*.
b. Alfred (λx (x thinks x is a great cook))
c. Alfred$_i$ (λx (x thinks he$_i$ is a great cook))

In (14b) the pronoun is a bound variable while in (14c) the pronoun represents an instance of coreference. Children confronted with sentences like *Grover touches him* have to assess whether the pronoun represents a bound variable or is coreferential in order to find out if the two noun phrases have an identical referent or if they have distinct referents.

One further explanation of the pronoun interpretation problem of English children is given by Grimshaw & Rosen (1990) who claim that if used

non-deictically, third person pronouns are difficult to interpret. Assuming that children do know that pronouns normally have discourse antecedents and that normally pronouns cannot be locally-bound, experimental test sentences of the sort *Annika is talking to her* will cause a conflict for the children. A non-target interpretation of *her* having the same referent as *Annika* may arise when children respect the pragmatics of pronouns but violate their syntactic requirements. Finally, Grodzinsky & Reinhart (1993) claim that for young children this task may prove to be too complex and thus they end up guessing.

Note that all of the solutions mentioned above hinge crucially either on reference to pragmatic information distinct from the syntax or on speculation about children's abilities to distinguish bound variable readings from coreference. None of these proposed solutions offers a straightforward syntactic explanation of why a supposedly innate syntactic principle should, in the acquisition phase, be systematically violated in the interpretation of anaphoric expressions but not in the production of these expressions. Below we suggest how the alternative to classical BT in section 3 can solve the pronoun interpretation problem without reference to pragmatics or processing difficulties.

As noted, the pronoun interpretation problem as described above presents an equally serious challenge to the alternative approach to binding phenomena advocated in section 3. To see why, consider a corpus like Table 5, i.e., one in which the Principle A and B pattern is strictly obeyed. We can use this corpus as a training corpus to simulate a grammar learned by a child learning modern English. Feeding BiGLA with twenty thousand pairs drawn at random based on the frequencies in Table 5 yielded the learning curves in Figure 5.

One can make the following observation: The constraints **pro,co* and **pro,dis* – the constraints which regulate the interpretation of pronouns – have distanced themselves from each other more quickly and to a greater degree than the constraints which regulate the interpretation of reflexives, **self,co* and **self,dis*. On the one hand, this is exactly what we should expect – because the vast majority of learning data were pronouns, the learner has learned the correct way to interpret of these expressions faster and more veraciously than he has learned to interpret the much rarer reflexive expressions. On the other hand, it contradicts what might be a commonsense intuition – that more common expressions might tend to be less restrictively interpreted – and, in fact, also contradicts the experimental data that constitute the pronoun interpretation problem. In this way, a frequency-based ap-

proach to binding phenomena is very seriously threatened by the pronoun interpretation problem, since it not only fails to predict that phenomenon but actually predicts exactly the opposite.

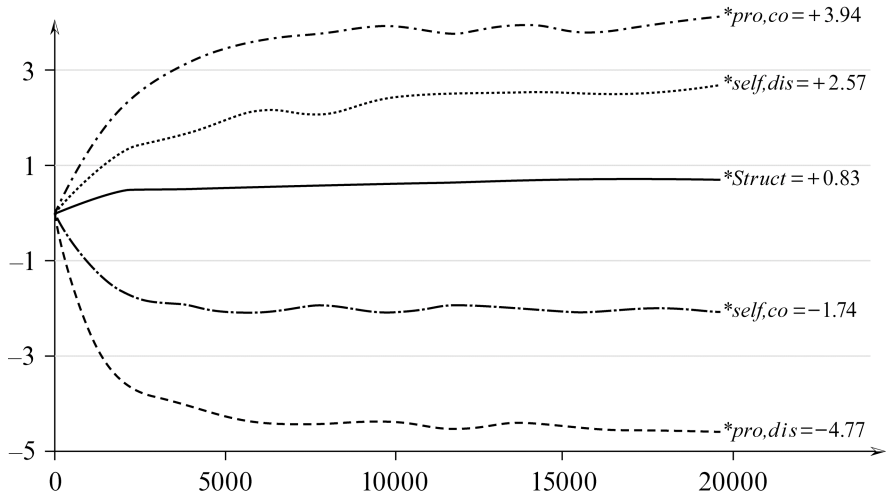

Figure 5. Learning curves (20K inputs) per Table 5

4.1. Addressing the pronoun interpretation problem

Below we present a solution to the pronoun interpretation problem.[16] The solution will hinge on an alternative definition of bidirectional optimality. In particular, we propose the following, revised definition.

(17) Revised bidirectional optimality

a. A meaning *m* is *recoverable* from a form *f* iff there is no form-meaning pair ⟨*f,m'*⟩ such that ⟨*f,m'*⟩ > ⟨*f,m*⟩.

b. A form-meaning pair ⟨*f,m*⟩ is *speaker optimal* iff either
   (i) *m* is recoverable from *f* and there is no pair ⟨*f',m*⟩ such that *m* is recoverable from *f'* and ⟨*f',m*⟩ > ⟨*f,m*⟩, or
   (ii) no form *x* is such that *m* is recoverable from *x* and there is no pair ⟨*f',m*⟩ such that ⟨*f',m*⟩ > ⟨*f,m*⟩.

c. A form-meaning pair ⟨*f,m*⟩ is *hearer optimal* iff there is no pair ⟨*f',m*⟩ such that ⟨*f',m*⟩ > ⟨*f,m*⟩.

The key difference in this definition can be found in (17c). Note that (17c) states that a meaning *m* is an optimal interpretation of a form *f* iff *f* is the optimal output for *m*, given the relevant, ranked set of (generative) constraints. In lay terms, when a hearer interprets an expression, he consults his own generative constraints and checks for which meaning that expression is optimal, ignoring bidirectional optimization. In other words, a language user is in effect assuming that his interlocutor possesses the same grammar he does, but while the language user himself employs blocking to ensure that each expression he generates is recoverable, he does not take for granted that his interlocutor does the same; he thus interprets an expression not according to interpretational constraints, per se, but according to what a fellow speaker would do if he wanted to express a certain meaning without respecting a recoverability restriction.

To see how this will make a significant difference when considering the learned grammar in Figure 5, let us consider the results after the first ten thousand inputs to the learning algorithm, shown in Figure 6 – this will more or less allow us to consider the grammar of a hypothetical six or seven year old child.

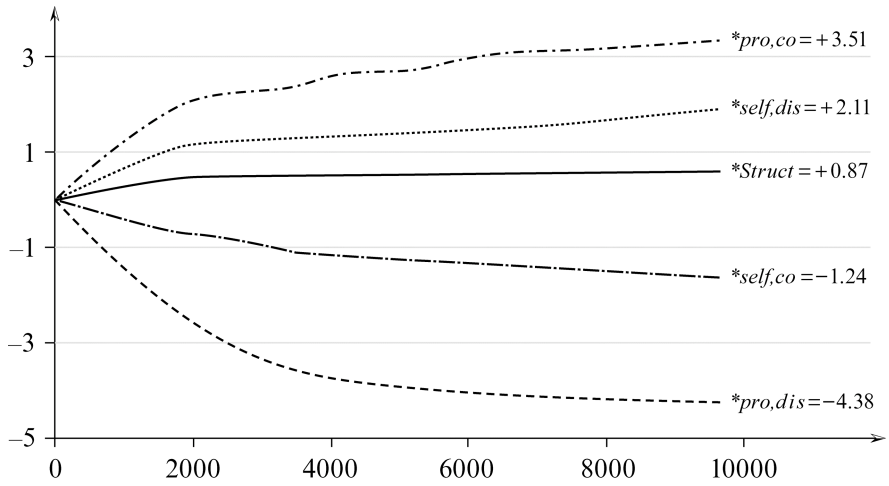

Figure 6. Learning curves (10K inputs) per Table 5

Let us consider first the generation of outputs for disjoint inputs, highlighted below in Figure 7.

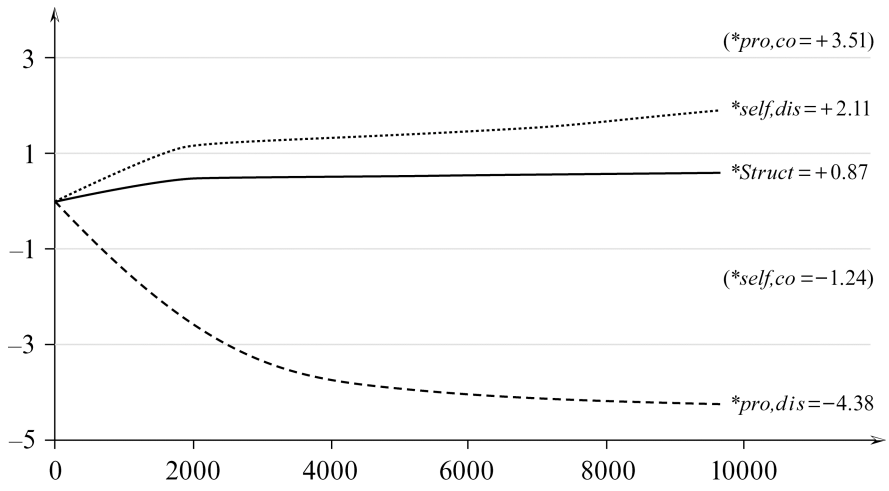

Figure 7. Learning curves relevant to disjoint predicates

From a generative perspective, the constraints **self,dis* and **Struct* will militate against *self*-marked outputs for disjoint inputs. The only competing constraint is the very low ranked **pro,dis*. The probability that **pro,dis* will outrank both **self,dis* and **Struct* for any given evaluation is far less than 1%.[17] This situation will insure that *self* -marked outputs will virtually never be used as outputs for disjoint inputs. Bidirectional optimization (i.e., blocking effects), which would also militate against using *self* -marked outputs for disjoint inputs (since the constraint **self,dis* significantly outranks **self,co* and thus *self*-marked outputs would not in general be recoverable as disjoint meanings), will rarely if ever be relevant, since the unidirectional, generative optimization will rarely if ever allow such a situation anyway. For this reason, from an interpretational perspective, per our revised definition of bidirectional optimality, namely (7c), a hearer will inevitably interpret a pronoun+*self* form as locally coreferential.

Consider now the generation of outputs for conjoint inputs, highlighted below in Figure 8.

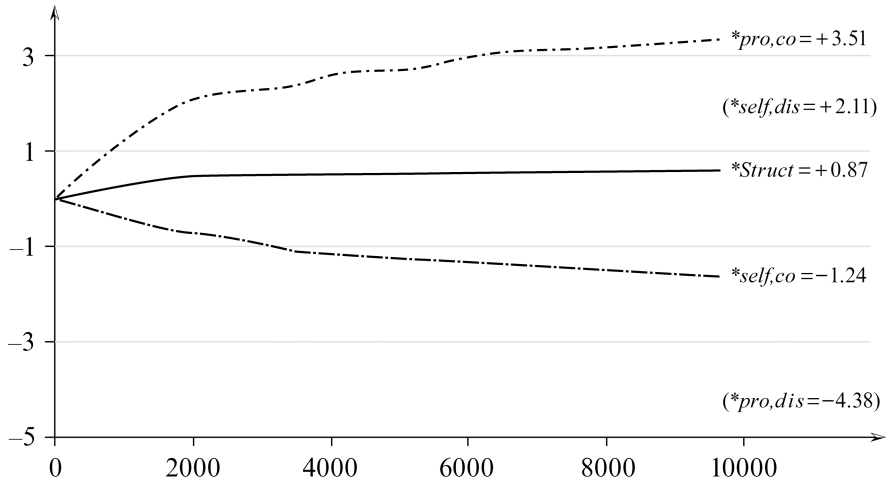

Figure 8. Learning curves relevant to conjoint predicates

The constraint **pro,co* is the dominant constraint, and will militate against using pronouns in situations where subject and object are coreferential. However, that constraint competes against two other constraints – **Struct* and **self,co* – which, while not as highly ranked, will 'gang up' against **pro,co*. In the case at hand, the odds that **self,co* will outrank **pro,co* are only about 5%, but the the odds that **Struct* will outrank **pro,co* are close to 20% and thus the odds that **pro,co* will outrank both of its competitors are only about 75%. Moreover, one can see from the trajectory of the learning curves in Figure 8 that the three constraints we are considering have grown steadily apart as the number of learning data has increased. At the point where, say, five thousand learning data had been absorbed, the odds that the generative constraints would favor a bare pronoun for a locally coreferential input would be close to 50%. From a production perspective, this will be irrelevant, since blocking effects will insure that bare pronouns are never used reflexively (since **pro,co* hugely outranks **pro,dis* and thus pronouns will, for all practical purposes, always be blocked as a means of soliciting a locally coreferential interpretation). On the other hand, from an interpretational perspective, per our revised definition of bidirectional optimality, namely (17c), this model predicts that a hearer will interpret a pronoun used by his interlocutor as locally coreferential somewhere between 25–50% of the time between the ages of, say, four and seven years of age. This prediction matches the experimental data quite nicely, and thus the pronoun interpretation problem is solved.

5. Conclusion

Above we have shown how a novel reformulation of bidirectional optimality can offer a way for a frequency-based, bidirectional theory of language acquisition to address the so-called 'pronoun interpretation problem' or 'delay of Principle B effect'. The problem was shown to be solvable when we view semantic interpretation as a strategy that calculates what a generative grammar would do in a particular situation (i.e., what output it would produce given some semantic input) and then interpreting an expression according to that calculation. If the pronoun interpretation problem is not unique to binding phenomena, but rather, as we suspect but have not proven, a more common phenomenon whereby more frequently used expressions are, in the acquisition phase, interpreted less restrictively rather than more restrictively, then the solution proposed above could offer prospects to any frequency-based and/or bidirectional analysis of language acquisition hoping to capture such a prediction.

Acknowledgements

Special thanks to Reinhard Blutner, for helpful discussion and for pointing out that the problem discussed in this paper needed to be solved. The sponsorship of the Zentrum für Allgemeine Sprachwissenschaft and Deutsche Forschungsgemeinschaft is also gratefully acknowledged.

Notes

1. An argument is 'locally bound' if it is c-commanded by an antecedent and co-indexed with it, but cf. Chomsky (1980, 1981, 1982, 1986) for a more detailed statement of classical BT. Also note that all the arguments below regarding the three shortcomings of standard BT are equally applicable to semantic reformulations of BT, e.g., that of Reinhart & Reuland (1991, 1993, 1995).
2. A notable exception to the monomorphemic pronouns/bimorphemic reflexives pattern are languages with so-called SE anaphora, cf., e.g., German *sich*. We forgo consideration of SE anaphora here, but cf. (Mattausch 2004, 2006) for discussion.
3. The cumulative distribution function is the probability that the variable X takes a value less than or equal to x, i.e., $F(x) = P(X \leq x)$.
4. Typically, it is assumed that the learner's grammar and his 'teacher's' grammar consist of the same set of constraints.

5. 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 happens at least some of the time. Cases where the observed meaning is not successfully recovered are ignored for the present purposes.
6. Currently, the software is available for download at no cost from *www.homes.uni-bielefeld.de/gjaeger/evolOT/index.html*. The x-axis in Figure 2 represents the number of form-meaning pairs being fed to the algorithm. The y-axis represents the ranking values of the various constraints. The simulation – and all the simulations in this paper – are conducted using evolOT, which is an implementation of the (Bi)GLA developed by Gerhard Jäger.
7. Or see Mattausch (2004, 2006) for further examples.
8. We are aware that many researchers in OT feel that the concept of 'markedness' is virtually meaningless and should be discarded, cf., e.g., (Haspelmath 2006). However, for our present purposes, the term 'markedness constraint' is synonymous with 'economy constraint' and thus is both well-motivated and harmless.
9. E.g., where V(C) is the ranking value of a constraint C: V(**Struct*) = V(**self,co*) + V(**self,dis*).
10. More specifically, for any set of ranked constraints C1 ≫ ... ≫ Cn, where ri is the ranking value of Ci and N is the standard normal distribution:
$P(C1 \gg ... \gg Cn) = \int_{-\infty}^{+\infty} dx\ 1N(x1 - r1) \int_{-\infty} x1\ dx2N(x2 - r2) \int_{-\infty} xn\text{-}1\ dxnN(xn - rn)$
Cf. Jäger (2003b) and Jäger & Rosenbach (2003) for more details.
11. The reader is referred to Mattausch (2004, 2006) for details.
12. The reader is once again referred to Mattausch (2004, 2006) for further details
13. The results discussed here were obtained in mismatch conditions. Chien and Wexler (1990) also tested match conditions which had less dramatic results.
14. Examples taken from Chien & Wexler (1990).
15. Examples adapted from Grodzinsky & Reinhart (1993).
16. It is only fair to note that an attempt at addressing the same problem within the framework of bidirectional OT has already been made by Hendriks & Spenader (2006). Forgoing any details, the analysis hinges crucially on the claim that a reflexive anaphor like himself is structurally(!) 'more economical' than a simple pronoun like *him* (2006: 12). The idea is scarcely defended there and, in our view, patently indefensible.
17. The actual calculations are left to the reader, cf. note 10, though precision in this regard is not at all crucial for the argument.

References

Blutner, Reinhard
2000 Some aspects of optimality in natural language interpretation. *Journal of Semantics* 17(3): 189–216.

Boersma, Paul
1998 *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D, diss., University of Amsterdam.
Boersma, Paul & Bruce Hayes
2001 Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.
Charney, Rosalind
1980 Speech roles and the development of personal pronouns. *Journal of Child Language* 7: 509–528.
Chiat, Shulamut
1981 Context-specificity and generalization in the acquisition of pronominal distinctions. *Journal of Child Language* 8: 75–91.
Chien, Yu-Chin & Kenneth Wexler
1990 Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition* 1 (3): 225–295.
Chomsky, Noam
1980 On binding. *Linguistic Inquiry* 11: 1–46.
1981 *Lectures on Government and Binding Theory*. Dordrecht: Foris.
1982 *Some Concepts and Consequences of the Theory of Government and Binding*. Cambridge: MIT Press.
1986 *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.
Farmer, Ann K. & Robert M. Harnish
1987 Communicative reference with pronouns. In *The Pragmatic Perspective*, Jef Verschueren & Marcella Bertuccelli-Papi (eds.), 547–565. Amsterdam: Benjamins.
Girouard, Pascale C., Marcelle Ricard & Thérèse Gouin Décarie
1997 The acquisition of personal pronouns in French-speaking and English-speaking children. *Journal of Child Language* 24: 311–326.
Grimshaw, Jan & Sarah T. Rosen
1990 Knowledge and Obedience: The Developmental Status of the Binding Theory. *Linguistic Inquiry* 21: 187–222.
Grodzinsky, Yosef & Tanya Reinhart
1993 The Innateness of Binding and Coreference. *Linguistic Inquiry* 24 (1): 69–101.
Horn, Laurence R.
1984 Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In *Meaning Form and Use in Context: Linguistic Applications*, Deborah Schiffrin (ed.), 11–42. Washington DC: Georgetown University Press.
Huang, Yan
1994 *The Syntax and Pragmatics of Anaphora: A Study with Special Reference to Chinese*. Cambridge: Cambridge University Press.

Huang, Yan
2000 *Anaphora: A Cross-linguistic Study*. Oxford: Oxford University Press.
Jäger, Gerhard
2003a Learning constraint sub-hierarchies: The Bidirectional Gradual Learning Algorithm. In *Optimality Theory and Pragmatics*, Reinhard Blutner & Henk Zeevat (eds,), 217–242. Houndmills: Palgrave MacMillan.
2003b Maximum entropy models and stochastic optimality theory. MS, University of Potsdam.
Jäger, Gerhard & Annette Rosenbach
2003 The winner takes it all – almost: Cumulativity in grammatical variation. MS, University of Potsdam and University of Düsseldorf.
Keenan, Edward L.
2000 An historical explanation of some binding theoretic facts in English. MS, www.linguistics.ucla.edu/people/keenan/historical.pdf, UCLA.
2001 Explaining the creation of reflexive pronouns in English. MS, www.linguistics.ucla.edu/people/keenan/shelpaper.pdf, UCLA.
Kirby, Simon & Jim Hurford
1997 The evolution of incremental learning: language, development and critical periods. Technical report, Language Evolution and Computation Research Unit, University of Edinburgh.
Levinson, Stephen
1991 Pragmatic reduction of the Binding Conditions revisited. *Journal of Linguistics* 27: 107–161.
2000 *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
Loveland, Katherine A.
1984 Learning about points of view: spatial perspective and the acquisition of 'I/you'. *Journal of Child Language* 11: 535–556.
Mattausch, Jason
2004 *On the Optimization and Grammaticalization of Anaphora*. Ph.D. diss., Humboldt Universität zu Berlin.
2006 Optimality, bidirectionality & the evolution of binding phenomena. In *Semantic Approaches to Binding Theory*, Alistair Butler, Ed Keenan & Jason Mattausch (eds.). Dordrecht: Kluwer.
Mitchell, Bruce
1985 *Old English Syntax, Vols. I–II*. Oxford: The Clarendon Press.
Reinhart, Tanya
1983 *Anaphora and Semantic Interpretation*. London: Croom Helm.
1986 Center and Periphery in the Acquisition of Anaphora. In *Studies in the Acquisition of Anaphora*, Volume 1, Barbara Lust (ed.), 123–150. Dordrecht: Reidel.

Reinhart, Tanya & Eric Reuland
1991 Anaphors and logophors: an argument perspective. In *Long-distance Anaphora*, Jan Koster & Eric Reuland (eds.), 165–174. Cambridge: Cambridge University Press.
1993 Reflexivity. *Linguistic Inquiry* 24: 657–720.
1995 Pronouns, anaphors and case. In *Studies in Comparative Germanic Syntax*. Hubert Haider, Susan Olsen & Susan Vikner (eds.), 241–269. Dordrecht: Kluwer.
Shannon, Claude
1948 A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379–432, 623–656.
Shipley, Elizabeth F. & Thomas E. Shipley
1969 Quaker children's use of thee: a relational analysis. *Journal of Verbal Learning and Verbal Behaviour* 8: 112–117.
Siemund, Peter
2000 *Intensifiers in English and German: A Comparison*. London: Routledge.
Visser, Frederikus T.
1963 *An Historical Syntax of the English Language*. Leiden: Brill.
Wilson, Colin
2001 Bidirectional optimization and the theory of anaphora. In *Optimality-theoretic Syntax*, Géraldine Legendre, Jane Grimshaw & Sten Vikner (eds.), 465–507. Cambridge, MA: MIT Press.
Zeevat, Henk
2001 The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics* 17: 243–262.