Forks in the Road to Rule I*

Irene Heim

Massachusetts Institute of Technology

1. Introduction

Tanya Reinhart pioneered and developed a new and very influential approach to the syntax and semantics of anaphora. It originated in Reinhart (1983a, b) and underwent various later modifications, e.g., Grodzinsky & Reinhart (1993), Heim (1993), Fox (1998, 2000), Reinhart (2000, 2006), Büring (2005). The central innovation concerned the architecture of the theory. The labor traditionally assigned to Binding Theory was broken up into two very different modules. One component (the "real" Binding Theory, if you will) regulates only one type of anaphoric relation, namely variable binding in the sense of logic. A new and different mechanism, variously thought of as a pragmatic principle, an economy constraint, and an interface rule, takes care of regulating other semantic relations, particularly coreference. The latter mechanism crucially involves the construction and comparison of alternative Logical Forms and their meanings.

I would like to reexamine the line of reasoning that has led to this bi-modular architecture. I will suggest that the problems it was meant to solve could have been addressed in a different way. My alternative proposal will borrow many essential moves from Reinhart, but her architectural innovation will effectively be undone.

2. Semantically Naive Binding Theory

The Binding Theory (BT) we teach in intro linguistics is built on observations about the possible readings of sentences like (1) and (2), and it takes the form of generalizations like those in (3).

(1) Bert pointed at him.

* Working on this paper has been a vivid reminder of how much inspiration and insight I gained from thinking about Tanya's work and from arguing with her when I was young. It also has reminded me of how much more I still can learn from thinking about her ideas today. We sadly cannot talk with each other any longer. But it is comforting to know that the many younger people whom Tanya also inspired and trained so well will surely not let me get away with anything either.

to appear in M. Abdurrahman, A. Schardl, M. Walkow (eds.), NELS 38 proceedings

(2) He was hungry when Bert arrived.

(3) B: A non-reflexive pronoun cannot corefer with a c-commanding DP in its local domain.
C: A non-pronominal DP cannot corefer with any c-commanding DP.

These principles are accompanied by appropriate definitions of "c-command" and "local domain", to which syntactians over the decades have devoted considerable sophistication. But what does "corefer" mean? The literal meaning would seem to be that α and β corefer iff there is an individual that both α and β refer to. But it is well known that for principles like (3B, C) to do their intended job, they can't invoke coreference in exactly this literal sense. If they did, they would rule out both too little and too much – too little because, e.g., they would not exclude the bound-variable reading in (4); and too much because they would not give an identity statement like (5) a chance to be true.

(4) Every boy pointed at him.
* 'every boy pointed at himself'

(5) This man is Stalin.

The task before us then, it seems, is to identify and define the appropriate coreference-like relation that should be plugged in where it says "corefer" in (3B, C); that is, to define it in such a way that these principles correctly draw the line between grammatical and ungrammatical readings. Or is that a hopeless task?

My reading of the field's history is that syntacticians up into the seventies tended to be optimistic about this. Postal (1970), for example, proposed that the correct notion was "presupposed coreference" and seemed to trust this notion could be made sufficiently precise in an appropriate semantic theory. But his discussion did not attend to bound variable anaphora at all. By around 1980, as awareness of the semantic heterogeneity of the relations targeted by BT had grown, the dominant attitude among syntacticians had changed. Binding Theory was part of syntax, and the relation it referred to was a syntactic concept, some formal property of syntactic representations. That this formal relation did not match up in a simple and straightforward way with a natural class of semantic relations was a curious fact, but "curious" more in the sense of "interesting" than of "problematic". This was just the typical situation: Natural language syntax has its own theoretical vocabulary and its concepts generally cannot be expected to match up one-to-one with concepts of semantics. Why should this hold any less for coindexing than for, say, grammatical function or case? People at the time also liked to point out that BT condition A lumps together reflexives with reciprocals, where surely the coindexing between a reciprocal and its antecedent had to mean yet something different than either coreference or variable binding.[1] Semanticists, of course, were welcome to apply themselves to working out the details of the mapping from indexed syntactic

[1] My main goal in Heim, Lasnik, and May (1991) was to respond to this point.

representations to meanings, but they should not expect it to be simple. BT generalized over a cluster of anaphoric relations, not because of what they had in common on a deeper semantic level, but because they were represented alike in the syntax.

Reinhart objected to this conclusion, but not to the pessimism it was based on in regard to the prospect of a semantic unification of coreference and variable-binding. In that respect, she agreed with the *zeitgeist* completely: binding and coreference were like apples and oranges in semantics. Only she went on to argue that they were like apples and oranges in syntax as well, they did not even have any common or similar syntactic representation. The superficial fact that bound-variable anaphora and coreference are prohibited in a lot of the same syntactic configurations was ultimately the result of a grand conspiracy, involving a syntactic Binding Theory that talked about variable-binding only and a very different module that pertained to other anaphoric relations.

But are we sure the pessimism wasn't premature? Maybe people in the eighties threw in the towel a bit too soon. What if we can, after all, give a non-disjunctive and empirically fitting definition of the broad "coreference"-like relation that BT conditions were originally meant to constrain? Let us use the hindsight and sophistication of the 21st century to give this another try.

3. Generalizing over Binding, Coreference, and Roundabout Codetermination

From here on I use Reinhart's term "covaluation" as the replacement for "coreference" in the formulation of BT, as in (6B, C). Our goal thus is a definition of this relation. We want to define "covalued with" so as to give (6B, C) just the right empirical coverage.

(6) B: A non-reflexive pronoun cannot be covalued with a c-commanding DP in its local domain.
C: A non-pronominal DP cannot be covalued with any c-commanding DP.

3.1 First Try

If it were just a matter of generalizing over variable-binding and coreference, i.e., ruling (1) and (4) out in one go, that would not be all that hard. Here is a first attempt:

(7) Two (occurrences of) DPs (of type e) are covalued in an utterance context iff they have the same extension under every variable assignment that extends the assignment given in that context.
I.e., α and β (occurrences of DPs of type e) are covalued in c (an utterance context) iff $[\![\alpha]\!]^{c,g} = [\![\beta]\!]^{c,g}$ for all $g \supseteq g_c$.

To make this definition work in the application to bound pronouns, we need to make an assumption about the syntactic representation of bound-variable anaphora, viz., that it is always dependent on movement of the antecedent (see Heim & Kratzer 1998). The LF of

the (ungrammatical) bound reading of (4) then is (8), and not, e.g., something less articulate like Büring's (9).

(8) every boy λ_1[t_1 pointed at him$_1$]

(9) every boy β_1[pointed at him$_1$]

In (8), BT-condition (6B) will apply to the pair of the subject-trace and the object-pronoun, which are covalued by definition (7). (The contextually given assignment here is empty, I assume, but whatever it is, coindexed variables always count as covalued.) (9) would not be ruled out, since it doesn't contain a pair of covalued DPs. In particular, the quantificational DP *every boy* isn't covalued with the pronoun (and wouldn't be even if we didn't restrict the definition to DPs of type e).

What is defined in (7) is arguably a fairly simple semantic concept, and the assumptions about LF-representation that need to go with it can be defended. But there still are problems, not conceptual ones but empirical ones. As had already been argued at some length by the mid-1980s, the present system still rules out both too little and too much. Let us first look at the "too much", which is easy to fix, and then at the "too little", which is more of a challenge.

3.2 Contingent Identity

Naive versions of conditions B and C were accused of excluding too much when it comes to identity statements and other sentences being used to offer indirect evidence for inferences about identity. This problem persists in the current proposal. It is sometimes argued that identity statements themselves may not be pertinent to the issue, since they may not contain referential (type-e) DPs in postcopular position in the first place.[2] It is therefore better to base the discussion on non-copular sentences that do not directly assert identity. Evans (1980) and Higginbotham (1980) had good examples involving potential condition C configurations, like Higginbotham's (10).

(10) (Was John the man in the bowler hat?) I don't know, but he put on John's coat before leaving, so it may well have been.

Analogous cases with pronouns in local (condition B) configurations are a little harder to set up, but they too exist. Here is one from Macià-Fàbrega (1997).

(11) I think that the man in the devil costume is Joe. It is suspicious that he knows him so well.

What is the problem? For naive binding conditions that directly legislate about coreference as in (3), the problem is that if the man in the devil costume is indeed Joe, then *he knows him well* under the intended reading is ungrammatical. Our current

[2] See e.g. Lasnik 1976, Macià-Fàbrega 1997.

version of BT in terms of the covaluation relation defined in (7) makes the same prediction. *he* and *him* are free pronouns, represented, say, as free variables he_i and him_j, where *i* may or may not be the same as *j*. Either way, if the contextually given assignment maps *i* to the man in the costume and *j* to Joe and if these two are the same individual, then every extension of the contextual assignment will map them to the same individual and they count as covalued. If one or both pronouns are treated as indexicals rather than free variables, they still are covalued if they corefer.

I endorse the same solution for this problem that I already recommended in 1993 and 1982, citing Postal (1970) as its earliest explicit source. The requirement that contexts map free indices (or indexicals) to referents must be rethought. There is a sense in which they do and a sense in which they don't, related to an equivocation in what we mean by a "context". There is what we might call the "objective" context of an utterance, which corresponds to the particular situation in which the utterance actually occurs, with all its properties both known and unknown to the discourse participants. The objective context does supply a specific individual for each indexical (directly) and for each free variable (indirectly via an assignment). Then there is the "subjective" context, which is a set of candidates for the objective context, the set of all the possible objective contexts that the utterance might be located in for all that the discourse participants presuppose. This is the notion familiar from the work of Stalnaker (1978), his "context set" or "common ground". A subjective context furnishes a set of possibly different assignments to the free variables in the utterance (or, in an alternative technical implementation, a unique assignment whose values are individual concepts rather than individuals). Let's use capital 'C' to stand for a subjective context, construed in this way as a set of possible objective contexts, each of which comes with its own contextually determined variable assignment. We want to redefine "covaluation" in terms of this notion.

(12) α and β (occurrences of DPs of type e) are covalued w.r.t. C (a subjective utterance context) iff for all $<w,g> \in C$ and all $g' \supseteq g$, $[\![\alpha]\!]^{w,g'} = [\![\beta]\!]^{w,g'}$.

The key idea is that two DPs might refer to the same individual in the actual objective context of the utterance, but not in every possible context that is a candidate for actuality according to the presuppositions of the interlocutors. Contexts in which the identity of Joe and the man in the costume can be felicitously asserted, denied, questioned, or supported or discredited by indirect evidence are *ipso facto* contexts in which the DPs *the man in the costume* and *Joe*, as well as any pronouns anaphoric respectively to the former and the latter, are not yet presupposed to corefer. They have the same extension in some contexts in the given subjective context and different extensions in some others. They therefore are not covalued in the subjective context according to the revised definition in (12), and (6B, C) will not rule out these examples.

By the way, this approach can remain agnostic about which, if any, referential DPs are treated as free variables. If, as Reinhart for example assumed, free pronouns are simply indexicals whose extension is directly fixed as a function of the context rather than through the detour of an index and a variable assignment, this will not make any

difference to how covaluation is assessed. Likewise, it doesn't matter whether we take proper names, complex demonstratives, and definite descriptions to contain a referential index and essential free variable or we treat them in more standard ways.

The moral here is that we must employ a notion of covaluation that is context-dependent in the right kind of way. Already in the first definition in (7), "covalued with" was relative to a context, so as to make it appropriately sensitive to the context-dependency of reference that is characteristic of most individual-denoting expressions. Here we took the further step to make it relative to a subjective notion of context, like Stalnaker's context set. This takes care of the complaint that the system wrongly ruled out true identity statements and certain other statements that in fact appear felicitously in deliberations about identity.

3.3 Codetermination via Third Parties

The system so far is designed to thwart both binding and coreference in condition B and C configurations, but it is unable to rule out certain sneaky derivations in more complex structures. This problem was, to my knowledge, discovered independently by Partee & Bach (1981) and Higginbotham (1983). Consider (13) with the derivation in (14).

(13) Every man said he knew that he pointed at him.

(14) underlying: every man said he_1 knew he_1 pointed at him_2
move subjects: every man λ_1. t_1 said he_1 λ_2. t_2 knew he_1 λ_3. t_3 pointed at him_2

The LF at the end of (14) contains no violation of condition B (6B). The two bound pronouns in the lowest clause are not covalued by our definition, since they are different bound variables. Neither is in the domain of the contextually given assignment(s), so we must consider all possible assignments to them, and among those there are plenty which assign them distinct values. But (14) means exactly the same as the LF in (15).

(15) *every man λ_1. t_1 said he_1 λ_2. t_2 knew he_2 λ_3. t_3 pointed at him_3

While (15) and some other equivalent LFs are properly ruled out by (6B), this doesn't do us much good if the same reading has another grammatical derivation. The English sentence in (13) cannot mean that every man said that he knew that he pointed at himself. We need to rule out all the ways to generate this reading.

Even though this problem was discussed in the early eighties by Higginbotham and Partee & Bach, its central significance to the syntax and semantics of BT was a little slow to be appreciated, perhaps because these authors discussed the problem in their own non-mainstream frameworks and thereby made it easy to overlook how hard it is to avoid it in just about any framework. As it turns out, it is this problem rather than any other shortcoming of naive BT that requires a major shift of strategy in the quest for an effective definition of covaluation. Given the existence of sneaky derivations like (14),

there is no way we can always detect covaluation within the local domain that contains the relevant pair of DPs. We need a definition that is somehow sensitive to the interpretation of the complete sentence in which the two potentially covalued DPs occur. But this does not mean that it cannot be done. In fact, if we have been paying attention to Reinhart's work and the Reinhart-inspired recent literature, it is easy.

(16) Let α and β be occurrences of DPs of type e in an LF ϕ, and let C be a subjective context. Then β is covalued with α in ϕ and C iff for all $<w,g> \in C$ and all $g' \supseteq g$, $[\![\phi]\!]^{w,g'} = [\![\phi\alpha/\beta]\!]^{w,g'}$, where $\phi\alpha/\beta$ is the result of replacing β by a copy of α in ϕ.

The idea is simple. A DP is covalued with another one if you could have repeated this other one in its place and still said the same thing.

Let us see how this applies. In a simple case of coereference in violation of condition B, like (1) *Bert pointed at him*, *him* is covalued with *Bert* (however exactly these two may be represented at LF) because *Bert pointed at Bert* would (in this context) mean the same thing. Note that the substitution instances constructed in the application of the definition do not need to be well-formed LFs themselves. They will often be BT violations like in this example. We just need them to be interpretable structures. Notice also that the new definition still is relativized to subjective contexts in the same way as the previous one, so we correctly predict that actual coreference is permitted, even in (1), if the pronoun is contextually associated with an individual concept that doesn't necessarily pick out Bert throughout the context set. (In that case, substituting *Bert* for the pronoun would have changed the meaning.) A simple case of BT-violating bound anaphora, as in (4) *Every boy pointed at him*, is also straightforward. As before, we assume that the bound reading must have an LF with a coindexed pronoun and subject-trace as in (8). There, him_1 and t_1 are covalued because *every boy* λ_1. t_1 *pointed at* t_1 means the same as *every boy* λ_1. t_1 *pointed at* him_1. This equivalence even is independent of context. (Again, we don't care that the structure with the two traces isn't part of a syntactically well-formed derivation.) Finally, take our problem case in (13). We can show now that even in the LF (14), the object pronoun him_2 in the lowest clause is covalued with its local subject, t_3, despite contraindexing. It is so because the result of substitution in (17) below is equivalent to (14).

(17) every man λ_1. t_1 said he$_1$ λ_2. t_2 knew he$_1$ λ_3. t_3 pointed at t_3

This then is the suggestion I want us to think about here. Old-fashioned Binding Theory may have been on the right track, after all, in its intent to apply indiscriminately to (presupposed) coreference, co-binding of variables, and maybe other more roundabout anaphoric relations. The project of pairing BT with the right sufficiently general notion of covaluation need not have been abandoned so quickly. We had to shed certain blinders, however, and stop insisting that the relevant relation had to be context-independent in either the pragmatic sense of "context" or the syntactic sense of "context". It turned out to be a relation that is crucially dependent on both the subjective utterance context and the global linguistic environment.

Now how exactly is this proposal different from the Reinhart-family of current approaches that I mentioned in the beginning? How in particular does it compare to Reinhart's own proposal in her last book, and to the Fox-Büring approach? These are questions I will speak to tentatively in the remainder of this paper.

4 Coverage of Data: Focus Constructions, Dahl Puzzle

4.1 Focus

I so far have ignored another type of example (besides identity sentences) that was supposed to show that naive BT sometimes ruled out more than it should. These are examples involving *only* and other focus particles, and the argument goes back to Reinhart's original work in the eighties.

(18) Despite the big fuss over Max's candidacy, in the end only he himself voted for him.

(19) Every devil fears that only he himself loves him.

In these examples, as Reinhart has taught us to view them, a naive condition B is exceptionally violated. The pronoun *him* in (18) can be read as referring to Max and hence as coreferring with its local subject (*he himself*); similarly, *him* in (19) can be the same variable as *he himself*, both bound by the quantifier *every devil*. What does our present system predict? Is *him* covalued with *he himself* in these examples on their relevant readings? Yes, it is. The LF for (19) on its intended reading, for example, is (20), and the result of substituting one instance of the same bound variable for another, as in (21), cannot fail to preserve meaning.

(20) every devil λ_1. t_1 fears that [only [he himself]$_1$] λ_2. t_2 loves him$_1$

(21) every devil λ_1. t_1 fears that [only [he himself]$_1$] λ_2. t_2 loves [he himself]$_1$

Does this mean we have found a problem with the present system and a disadvantage that it has against the Reinhart-style theories on the market?

I want to argue that no, we don't have a problem – but moreover and more importantly, that these examples are simply beside the point in the present discussion. In fact, they were beside the point all along, even in previous discussions where they were used to bolster Reinhart's approach. The key point is that we have no good reason to assume that the two covalued DPs in (20) stand in a syntactic condition B configuration in the first place, i.e., that one c-commands the other. We must distinguish between the larger DP *only he himself* and the smaller DP embedded therein which is just *he himself*. Only the former clearly c-commands into the VP, the latter maybe does not. If there is no c-command, then the two DPs can be covalued all they want, just like any other pair of an object and a pronoun embedded inside the subject.

(22) [his$_1$ mother] λ_2. t_2 voted for him$_1$

Büring (2005) considers this approach in passing and deems it a non-starter because the relevant examples also come in a variant where *only* does not attach to the subject. For example, (19) can be changed to (23).

(23) Every devil only knows that he himself loves him.

The intended reading here is that every devil x knows that x loves x and doesn't know for any y ≠ x that y loves x. The higher scope of *only* makes a difference to the truth conditions, but the point is still that there is a reading on which the two pronouns are cobound. I respond that the LF-representation of this reading looks as in (24), and crucially does not look as in (25).

(24) every devil λ_1. t_1 only knows that [[he himself]$_1$ F] λ_2. t_2 loves him$_1$

(25) every devil λ_1. t_1 only knows that [[he himself] F]$_1$ λ_2. t_2 loves him$_1$

The *F* represents the focus that *only* associates with. I take it to be its own node in the syntactic structure, a sister to the phrase it F-marks, and sufficient to disrupt the c-command relation just as much as *only* in (20). This highly articulated structure may be unaccustomed, but it makes perfect sense and in fact is unavoidable in a standardly compositional semantics. If you look at the interpretation for F-marking in Rooth (1985), you clearly see that the interpretation of an expression of the form α_F is defined (by a syncategorematic rule, as it happens there) as a function of the interpretation of α. The two thus cannot be the same phrase, but one must properly contain the other. You also see that (25) is uninterpretable. For the semantic computation to get off the ground, *F* needs an interpretable sister, which the unindexed pronoun here is not.[3] And the index outside the *F* in (25) is not interpretable in that position.

In sum, I maintain that Büring's variation on the examples does not really make a difference. Wherever the *only* itself may be attached, the Focus that it associates with creates a layering of the focussed DP, and therefore the covaluation that is permitted in alleged violation of our condition B is not between a pair of DPs in a c-command configuration. If I am on the right track here, we ought to find that even in the absence of overt exclusive or additive particles (*only*, *even*), the presence of focus on the antecedent should suffice to license the apparent exemption from condition B. I think this is quite defensible. E.g., we find the same acceptability status in question-answer sequences.

(26) Who voted for Max?
He HIMSELF voted for him.

[3] At least not in the intended bound-variable reading. Maybe unindexed pronouns are interpretable as referential terms, but this is not relevant here.

We might also look at Evans's famous examples, or their condition-B variants that I grouped under "structured meaning cases" in Heim (1993), as instances of the same phenomenon.

(27) You know what Mary, Sue and John have in common? Mary admires John, Sue admires him, and John admires him too.

(28) Look, if everyone hates Oscar, then it surely follows that Oscar himself hates him.

These are pronounced with contrastive focus on the relevant subjects.[4]

It is sometimes remarked that the covalued readings of examples like (18) and (19) have a sort of intermediate grammaticality status, not as impossible as bound readings but also not as good as coreference in the uncontroversial absence of c-command, as in (22). This complicates the data picture, but what does it show? Perhaps there is some squishiness to the relevant notion of c-command that matters to BT, with certain configurations qualifying as c-command for some speakers on some occasions and not on others. Or, perhaps more plausibly, the human processor can detect the presence or absence of c-command more quickly and effortlessly in some configurations than in others. Maybe when c-command is interrupted not by lexical material but by merely functional or even non-segmental items, this is harder to detect and we garden-path. This type of explanation should not be dismissed out of hand.[5]

4.2 Dahl's "Many Pronouns" Puzzle

Danny Fox (1998, 2000) connected a new set of facts to the discussion about the proper approach to Binding Theory: a certain pattern of strict and sloppy readings in VP ellipsis that was first observed by Dahl.

(29) Max said he liked his paper, and Lucie did too.
'Max said Max liked Max's paper, and
(a) ... Lucie said Lucie liked Lucie's paper.' (sloppy-sloppy)
(b) ... Lucie said Max liked Max's paper.' (strict-strict)
(c) ... Lucie said Lucie liked Max's paper.' (sloppy-strict)
(d) *... Lucie said Max liked Lucie's paper.' (*strict-sloppy)

[4] But see Grodzinsky (2007) for a recent reassessment of the data. Grodzinsky argues, in effect, that examples like (27), (28) are only felicitous in contexts where the two supposedly covalued DPs actually are associated with distinct individual concepts. If that turns out to be right, then I just need to say that there is c-command, after all. I can remain agnostic about this here.

[5] Schlenker (2005) argues that examples like (18) are not really grammatical, except in the irrelevant case where the same individual is picked out under two different guises. This is the same type of claim as Grodzinsky's, except applied to a different class of data. If Schlenker turns out to be right, it is fine with me too.

The theory I have presented here has nothing to say about this. It is a theory only about how BT conditions B and C apply, and this example does not involve any pairs of DPs to which these conditions are relevant. The pronouns *he* and *his* in the antecedent sentence can each be bound by or coreferential with the matrix subject, and if this is so, then (by all accounts of the licensing conditions on ellipsis) the continuation should allow all four of the listed readings.

Let me make two remarks that put this result in perspective, even if they don't make it acceptable. First, as Fox concedes frankly, Büring acknowledges, and Reinhart (2000, 2006) and Roelofsen (2007) emphasize, Fox's account of Dahl's puzzle relies on a specific stipulation about parallelism in ellipsis that is not deducible (to our current knowledge) from a general theory of ellipsis licensing.[6] This casts doubt on whether it is the right account, even in the absence of a working alternative. Second, Roelofsen (2007) has shown that Reinhart's alternative account of Dahl's puzzle, as presented in Reinhart (2000, 2006), does not work as it stands. The reasoning she applies to rule out the strict-sloppy reading will *mutatis mutandis* rule out the sloppy-strict reading as well. (Reinhart's reasoning is complicated, and I refer to Roelofsen's paper for details.) This means that Fox's account is the only working account that links the absence of Dahl's strict-sloppy reading to the ungrammaticality of covaluation in BT configurations.[7] Perhaps Dahl's data then are a decisive reason to abandon the approach I am exploring in this paper and to favor the Fox-Büring approach instead. I am not prepared to insist that they aren't. It is, however, worth pointing out how much rides on the Dahl puzzle if this is so. The discussion in Fox and Büring did not convey this impression. It rather suggested that the solution to this puzzle came as an added benefit to an independently motivated solution to the problems of naive BT.

5. Syntactic Principles, Interface Strategies, and Psycholinguistic Evidence

Reinhart's late work on Binding Theory was part of a much broader project. In her 2006 book, the overarching theme is the division of labor between two cognitive systems, a Computational System (CS) or syntax proper that performs only local computations and a set of Interface Strategies that involves costlier operations which compare alternative derivations. Her analysis of constraints on anaphora was meant to provide just one of a series of examples of this division. There was a simple condition B that governed only variable-binding relations in local domains, which was part of syntax (CS), and then there was Rule I, which constructed reference sets of alternative LFs and compared their meanings in context, which operated at the Interface. What will happen to this picture if we return to a Binding Theory which, as I am entertaining here, regulates covaluation relations of all semantic stripes all at once and in a uniform way? Where does this put my principles B and C within Reinhart's overall architecture? The answer would seem to be that BT altogether would then reside at the Interface, since it always, in all its

[6] Schlenker (2005) also relies on assumptions about ellipsis that have not so far been embedded in a principled theory of ellipsis licensing, I think.

[7] Roelofsen (2007) also develops an account of Dahl's puzzle in terms of a Fox-like locality constraint on binding, but he argues (for reasons of his own) that this constraint is independent from what rules out covaluation in BT configurations.

applications, depends on the relation of covaluation. Covaluation cannot be read off from local configurations by means of elementary syntactic or semantic computations, it can only be detected by constructing alternative LFs for whole utterances and comparing global, context-dependent meanings. So a principle that mentions covaluation is *ipso facto* outside of syntax and belongs at the Interface. Could this be right?

Reinhart points to psycholinguistic evidence, particularly from language acquisition, in support of her separation between syntactic binding principles and interface strategies. Chien & Wexler (1990) and many subsequent experimental studies found that young children around age 5 do not reliably reject readings that involve coreference in violation of condition B, but they do reliably reject bound-variable readings in the same structures. Reinhart interprets this as a sign that there is an "easy" part of conventional BT that these young children master like adults and a "difficult" part which they don't. The easy part is the syntax-internal principle B that regulates bound-variable anaphora, and the difficult part is the application of the Interface Rule I to potential coreference readings. Children do not always succeed in applying Rule I because of their limited short-term memories. More precisely, Reinhart argues, <u>some</u> applications of Rule I are hard. Many are easy, even for young kids, namely those where you only need to go to clause (a) or (b) to obtain a verdict. It gets hard only when you have to go all the way to check the final clause (c), which is the one that requires computing and comparing two meanings.

(30) Rule I from Reinhart (2006):
α and β cannot be covalued in a derivation D, if
a. α is in a configuration to A-bind β, and
b. α cannot A-bind β in D, and
c. the covaluation interpretation is indistinguishable from what would be obtained if α A-binds β.

Now if all of BT belonged to the Interface, with no part left within syntax, what would that imply about easy and hard tasks and the performance of children with limited short term memory? One might think at first that everything should then be hard. Every application of condition B requires checking covaluation, which in turn requires constructing a full alternative LF and computing and comparing its meaning. Even ruling out local bound-variable anaphora in *Every bear is touching her* should be hard and we should see little kids failing on it. But this doesn't really follow. The processor might well be equipped with strategies that make certain applications of BT a lot easier than others. For example, it is a simple lemma of our definition of covaluation that two coindexed DPs are always covalued. One doesn't need to interpret the whole bigger structure in each case to see this. Any processor worth its money will be designed to take advantage of this fact. The definition of covaluation is such that in the worst possible case, you must compute the meanings of the entire two structures to see whether a pair of DPs is covalued. But the worst case does not always obtain and many times it suffices to examine the meanings of smaller domains. In fact, we have this useful theorem:

(31) Let α and β be DP-occurrences in ψ, which in turn is embedded in ϕ. Then if β is covalued with α in ψ and C, β is also covalued with α in ϕ and C.

If covaluation obtains with respect to a smaller domain, it obtains in every larger domain as well. Only the opposite isn't guaranteed: we may have two DPs that are not covalued in a local domain, but are covalued in the structure as a whole. (This was the situation in (14).)

The upshot of these remarks is that in the absence of further assumptions about the human processor, we don't know much about how costly it would be to apply the Binding Theory, even if all of it consisted of Interface rules in the technical sense of Reinhart's distinction. We don't know, in particular, which BT-related tasks would be hard for young children.[8]

6. Covaluation and Syntactic Representation

So far I have contemplated a theory in which Binding Theory employs a notion of covaluation which is global and context-dependent. No local and/or context-independent relation between two DPs or their meanings seems to be a possible candidate for "covaluation" if this relation is to cover all the different cases in which conditions B and C are at work. But there is a strategy that we can use to hold out against this conclusion. We can make more substantive assumptions about possible syntactic derivations and about the representation and semantic interpretation of DPs. In a nutshell, if we sufficiently constrain the syntax and semantics of the LF language, we can make it so that two DPs are never covalued unless they are coindexed. In that event, we can formulate BT in terms of coindexing as in the 1980s and readmit it into ordinary syntax. Let us see how this might go.

The Binding conditions would then read as follows.

(32) B: A non-reflexive pronoun cannot be coindexed with a c-commanding DP in its local domain.
C: A non-pronominal DP cannot be coindexed with any c-commanding DP.

Indices, as before, have a semantics as variables, and syntax and semantics conspire in such a way that the ungrammatical bound interpretation of (4) is only associated with an LF as in (8), which appropriately violates (32B). But some not-so-standard assumptions must be added to ensure that the prohibitions against coindexing rule out not just binding but also coreference. First of all, it must be guaranteed that all type-e DPs are generated with an index and that this index enters into their interpretation. Otherwise, DPs without indices could so to speak fly under the radar of BT and refer to whatever they please.

[8] I should add in this connection that several recent authors (Elbourne 2005, Grolla 2005, Conroy et al. 2007) have suggested alternative causes for children's differential performance on bound-variable and coreference anaphora in the experiments by Chien & Wexler and others that Reinhart relies on. This field is currently quite open.

Roughly, this requires that proper names and complex definites receive a presuppositional semantics along the lines of (33).

(33) $[\![\text{John}_i]\!]^g = g(i)$ if $g(i) = \text{John}$, otherwise undefined
$[\![\text{the boy}_i]\!]^g = g(i)$ if $g(i)$ is a boy, otherwise undefined

Unindexed *John* or *the boy* must be either underivable or uninterpretable, so we can't have an LF like (34) with a contextually given variable assignment as indicated, which would express coreference without being filtered out by (32B).

(34) Bert pointed at him_1
$g_c = [1 \rightarrow \text{Bert}]$

Moreover, it seems we must make sure that contextually given assignments are generally one-to-one, or else an LF-and-context pair like (35) could once again smuggle in coreference below the radar of (32B).

(35) Bert_1 pointed at him_2
$g_c = [1 \rightarrow \text{Bert}, 2 \rightarrow \text{Bert}]$

Given that we don't want to rule out coreference in identity statements, however, the one-to-one requirement should not be imposed on the assignment furnished by the objective context, but suitably relativized to the subjective context. If subjective contexts are construed as sets of world-assignment pairs as sketched above, the following formulation will work.

(36) A subjective context C can be paired with an LF ϕ only if
for every pair of distinct indices $i \neq j$ free in ϕ,
there is some $\langle w,g \rangle \in C$ such that $g(i) \neq g(j)$.

There are legitimate concerns about the stipulative and cumbersome nature of these non-standard assumptions. Things get complicated when we have definite DPs whose NPs contain other variables, and we need complex indices in this case (of the kind employed in work on functional questions, e.g., by Chierchia 1993). I will spare you these details. But even apart from technical complexity, it is disturbing to have to enforce the presence of indices on all referential DPs, including those that otherwise give us no reason to implicate a variable in their meaning: proper names, definite descriptions, demonstratives, and indexicals. Even in the domain of pronouns, the trend of recent semantic research has arguably been to remove more and more of these from the realm of variables and reclassify them instead as covert descriptions ("E-Type" pronouns) or as indexicals (see Kratzer 2006 for a particularly radical instance of this trend) – in other words, the opposite of what we are forced to do here. But this seems to be the price of enabling a BT which only sees a context-independent relation to handle the facts that earlier persuaded us of the context-dependency of the covaluation relation.

As if those concerns were not enough, we still have the problem that taught us that covaluation needed to be relativized not only to subjective context but moreover to the global syntactic environment. In other words, we are still vulnerable to the sneaky derivations exemplified by (14).

(14) every man λ_1. t_1 said he_1 λ_2. t_2 knew he_1 λ_3. t_3 pointed at him_2

Since there is no coindexing within the lowest clause, the coindexing-based condition B in (32) cannot rule this out. I considered a way of plugging this loophole in Heim (1993), which I dismissed at the time for a reason that may not, after all, be valid. As I there observed, (14) and all the other problematic derivations of this kind stand and fall with the possibility of choosing any index you like for the binder-index and trace when a DP is moved. For example, when deriving (14), I chose *2* for the lambda and trace of the higher moving pronoun, which was different from that pronoun's own index *1*. Had I been forced to reuse the pronoun's preexisting index *1* on the new lambda and trace, my devious plot would not have gone far.

How might we think about the syntax of movement so that it doesn't give us so much freedom in choosing indices? In Heim & Kratzer, where binder indices are created out of thin air in each move, it is hard to constrain it in a non-stipulative way. But in more recent work by Kratzer (2001), the picture is a little different. Binder indices are generated first, and movement is a matter of the binder index attracting a matching index on the moving DP. After this has happened, the DP's own index then may – but presumably doesn't have to – be deleted under agreement with the attracting binder index. A Kratzer-style derivation for a simple sentence like (37) proceeds as in (38).

(37) Every boy pointed at himself.

(38) underlying: λ_1. every boy_1 pointed at $himself_1$
move: every boy_1 λ_1. t_1 pointed at $himself_1$
delete uninterpretable index: every boy λ_1. t_1 pointed at $himself_1$

Here the moving DP was a quantifier, on which an index is not interpretable. But had it been a referential DP instead, which (as I am already assuming now) can and must bear an interpreted index at LF, we would naturally have stopped short of the final step.

(39) Bert pointed at himself.

(40) underlying: λ_1. $Bert_1$ pointed at $himself_1$
move: $Bert_1$ λ_1. t_1 pointed at $himself_1$

This way, whenever movement applies to a type-e DP, the same index will be on the moved DP and on its trace and any pronouns it binds. This prevents derivations like (14) and closes the loophole out of condition B.

In Heim (1993), I argued that this was not the right way to go, because plugging the loophole in this way would also reduce the system's expressive power in undesirable ways. My argument turned on strict-sloppy ambiguities in ellipsis, particularly the ambiguity of a sentence like (41).

(41) Every boy said that he was mad at his mom and John was too.
'every boy x said that x was mad at x's mom and ...
a. ... John was mad at x's mom' (strict)
b. ... John was mad at John's mom' (sloppy)

I assumed with Sag, Williams, and Reinhart that ellipsis licensing required semantic equivalence between the deleted and antecedent VPs. Therefore the emergence of two readings for the elided VP meant that there had to be two distinct representations for its antecedent. This in turn meant that it had to be possible to distinguish the index on the bound *he* from the index on the variables that it in turn binds. The sentence in (41) had to have two different LFs as in (42a,b).

(42) every boy λ_1. t_1 said that
a. [he_1 λ_2. t_2 was mad at his_1 mom] and [John λ_3. t_3 was mad at his_1 mom]
b. [he_1 λ_2. t_2 was mad at his_2 mom] and [John λ_3. t_3 was mad at his_3 mom]

What I just said above does not allow this anymore. The lambda next to he_1 must have the same index *1*, so there is only one single choice for indexing the *his* in the overt VP.

(43) every boy λ_1. t_1 said that
[he_1 λ_1. t_1 was mad at his_1 mother] and [$John_3$ λ_3. t_3 was mad at $his_?$ mother]

Can I still account for the ambiguity in (41)? I may be able to if I give up the restrictive theory of ellipsis-licensing that requires equivalent VPs. Were I to stick to that, the index on the elided $his_?$ in (43) would have to be *3* and the reading would have to be sloppy. But on looser theories of ellipsis-licensing such as those argued for by Rooth (1992), Fox, and others, we can have two choices for the elided pronoun even when we have only one for its overt counterpart. (Though as I already mentioned, there are serious concerns over the prospects of a principled theory of ellipsis that will allow just the right cases of strict readings of pronouns whose counterparts in the antecedent are locally bound.)

Constraining the LF-language in such a way that a moving DP must always leave a copy of its own index on its trace has a similar effect as Fox and Büring's Binding Economy requirement.[9] Fox and Büring assume (as I did in Heim 1993 and elsewhere, including earlier in this paper) that LF-syntax can in principle distinguish between "cobinding" as in (44a) and "transitive binding" as in (44b).

[9] I lump the two together here, even though, as Büring points out, Fox's statement is less general.

(44) a. every boy λ_1. t_1 said that he_1 λ_2. t_2 was mad at his_1 mother
b. every boy λ_1. t_1 said that he_1 λ_2. t_2 was mad at his_2 mother

But they impose an economy constraint (Fox's "Rule H", Büring's "Have Local Binding") which effectively filters out (44a).

(45) Binding Economy:
For any two DPs α and β, if α could semantically bind β (i.e., if it c-commands β and β is not semantically bound in α's c-command domain already), α must semantically bind β, unless that changes the interpretation.

(46) Definition of "semantic binding": α sem-binds β iff α is sister to a constituent of the form $[\lambda_i. \phi]$, where ϕ contains β_i and no instance of λ_i that c-commands β_i.

Let α and β in (44a) be respectively the occurrences of *he*$_1$ and *his*. C-command obtains and *his* is not sem-bound in the scope of *he*$_1$. So (45) demands that *he*$_1$ sem-bind *his*, which it doesn't in (44a) but does in the otherwise identical (44b), unless these two have different interpretations. Since they are in fact semantically equivalent, (45) disallows (44a). Only (44b) is allowed.

Binding Economy eliminates (44a) by making it compete with (44b), whereas the Kratzer-style syntax for movement that I am currently contemplating simply doesn't generate (44a) in the first place. It actually doesn't generate (44b) either, but it does generate (47), which exhibits an identical pattern of binding relations.

(47) every boy λ_1. t_1 said that he_1 λ_1. t_1 was mad at his_1 mother

Both proposals then, in different ways, predict that transitive binding is the only option.

Actually, this is not quite right, unless we make explicit another assumption, namely that the lower subject must undergo some movement (whether it ends up binding a pronoun or not). This assumption is not needed by Fox and Büring, since their principles are meant to apply in the same way to (48) below as to (44a), making both compete with and lose against (44b).

(48) every boy λ_1. t_1 said that he_1 was mad at his_1 mother

But a Kratzer-style derivation could in principle yield (48), provided it were possible to leave *he*$_1$ in its base-generated position and never introduce a lower lambda at all. The point is moot if there are independent reasons for subjects to have to move anyhow, which is certainly something widely assumed for full clauses in English. We would need to turn to other types of structures if we wanted to tease the two theories apart; perhaps certain kinds of small clauses, or languages where subjects may remain *in situ*, or cases

where the potential local binder is not a subject in the first place. Maybe even direct objects always move, but what about a prepositional object as in (49)?

(49) John said I talked to him about his grades.

Might this be a case where the present proposal would differ from Binding Economy, in allowing the derivation of a cobinding constellation that Binding Economy would block? I don't know.

If we could tease the theories apart, where would it show in the data? Fox uses Binding Economy to attack Dahl's "many pronouns" puzzle. Insofar as the Kratzer-style movement syntax enforces the same transitive-binding constellations, his solution would carry over (albeit with the same baggage of a parallelism condition that remains *ad hoc*). If there are differences between the precise ranges of constructions in which the two theories prevent co-binding, we could get slightly different predictions about the distribution of the Dahl-effect.

Let me step back to reflect on this section. We had found in the foregoing discussion that covaluation, the relation that seems to matter to Binding Theory, was not detectable locally and context-independently. At least this was so under more or less standard assumptions about syntactic representations and compositional semantics. But what if these assumptions are not right? It certainly remains conceivable that grammar is, after all, designed in such a way that the global and context-dependent relation of covaluation can be read off reliably and easily from local chunks of representation. I outlined some maneuvers that we might make to create this state of affairs. As I am the first to concede, these particular maneuvers were inept and unexciting (except possibly for the last one, which also had an interesting side-effect). Probably this is not quite the right way to go. But the general possibility that covaluation has some easy-to-spot representational correlate is worth bearing in mind.

7. Conclusion

This paper had two parts. In the first part I proposed a uni-modular approach to Binding Theory, in which BT conditions regulate a semantically broad relation of "covaluation", which includes both binding and coreference. The definition of this relation makes reference to alternative representations and their meanings, and in this respect is reminiscent of the mechanisms employed by Reinhart and her descendants in their interface rules and economy principles. This proposal raises new questions about the role of BT in language processing and acquisition, which I have not yet even begun to explore. It also raises a question about the explanation of Dahl's puzzle. The second part of the paper was not meant as a proposal, but should be taken more as merely an existence proof. My main point there is that the basic proposal in the earlier part does not in itself say all that much about how anaphoric relations are represented in the syntax, and that different imaginable answers to this further question will affect the content of any predictions about processing, acquisition, or the Dahl puzzle.

References

Büring, D. 2005. Bound to bind. *Linguistic Inquiry* 36(2), 259–274.

Chien, Y.-Ch. and K. Wexler. 1990. Children's knowledge of locality conditions in binding as evidence of the modularity of syntax and pragmatics. *Language Acquisition* 1: 225–295.

Chierchia, G. 1993. Questions with quantifiers. *Natural Language Semantics* 1(2): 181–230 .

Conroy, A., E. Takahashi, J. Lidz, and C. Phillips. 2007. Equal Treatment for all antecedents: How children succeed with principle B. Ms. University of Maryland.

Elbourne, P. 2005. On the acquisition of principle B. *Linguistic Inquiry* 36(2): 333–365.

Evans, G. 1980. Pronouns. *Linguistic Inquiry* 11(2): 337–362.

Fox, D. 1998. Locality in variable binding. In P. Barbosa et al., eds., *Is the Best Good Enough? Optimality and Competition in Syntax*, 129–155. Cambridge Mass.: MIT Press.

Fox, D. 2000. *Economy and Semantic Interpretation*. Cambridge Mass.: MIT Press.

Grodzinsky, Y. 2007. Coreference and self-ascription. Ms. McGill University, Montreal.

Grodzinsky, Y. and T. Reinhart. 1993. The innateness of binding and coreference. *Linguistics Inquiry* 24(1): 69–101.

Grolla, E. 2005. *Pronouns as Elsewhere Elements: Implications for Language Acquisition*. Doctoral dissertation, University of Connecticut, Storrs.

Heim, I. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Doctoral dissertation, University of Massachusetts, Amherst. (Published in 1988 by Garland, New York.)

Heim, I. 1993. Anaphora and semantic interpretation: A reinterpretation of Reinhart's approach. SfS-Report 07-93, University of Tübingen. (Reprinted in 1998 in U. Sauerland and O. Percus, eds., *The Interpretive Tract*, 205–246. MITWPL, Department of Linguistics and Philosophy, MIT.)

Heim, I. and A. Kratzer. 1998. *Semantics in Generative Grammar*. Oxford: Blackwell.

Heim, I., H. Lasnik, and R. May. 1991. Reciprocity and plurality. *Linguistic Inquiry* 22(1): 63–101.

Higginbotham, J. 1980. Anaphora and GB: Some preliminary remarks. In J. Jensen, ed., *Cahiers Linguistiques d'Ottawa: Proceedings of NELS 10*, University of Ottawa.

Higginbotham, J. 1983. Logical form, binding, and nominals. *Linguistic Inquiry* 14(3): 395–420.

Kratzer, A. 2001. The event argument. Ms. Semantics Archive.

Kratzer, A. 2006. Minimal Pronouns. Ms. Semantics Archive.

Lasnik, H. 1976. Remarks on Coreference. *Linguistic Analysis* 2(1): 1–22.

Macià-Fàbrega, J. 1997. *Natural Language and Formal Languages*. Doctoral dissertation, MIT.

Partee, B. and E. Bach. 1981. Quantification, pronouns, and VP anaphora. In *Formal Methods in the Study of Language*. Mathematical Centre tracts, University of Amsterdam. (Reprinted in 1984 in J. Groenendijk et al., eds., *Truth, Interpretation and Information*, 99–130, Dordrecht: Foris.)

Postal, P. 1970. On coreferential complement subject deletion. *Linguistic Inquiry* 1: 349–500.
Reinhart, T. 1983a. *Anaphora and Semantic Interpretation*. Chicago: University of Chicago Press.
Reinhart, T. 1983b. Coreference and Bound Anaphora: A Restatement of the Anaphora Questions. *Linguistics and Philosophy* 6(1): 47–88.
Reinhart, T. 2000. Strategies of Anaphora Resolution. In Hans Bennis, M. Everaert, and E. Reuland, eds., *Interface Strategies*, 295–324. Amsterdam: Royal Netherlands Academy of Arts and Sciences.
Reinhart, T. 2006. *Interface Strategies: Optimal and Costly Computations*. Cambridge: MIT Press.
Roelofsen, F. 2007. Bound and referential pronouns. Ms. University of Amsterdam, Institute for Logic, Language, and Computation.
Rooth, M. 1985. *Association with Focus*. Doctoral dissertation, University of Massachusetts, Amherst.
Rooth, M. 1992. Ellipsis redunancy and reduction redundancy. In S. Berman and A. Hestvik, eds., *Proceedings form the Stuttgart Ellipsis Workshop*. Institut für maschinelle Sprachverarbeitung, University of Stuttgart.
Schlenker, P. 2005. Non-redundancy: Towards a semantic reinterpretation of binding theory. *Natural Language Semantics* 13(1), 1–92.
Stalnaker, R. 1978. Assertion. In P. Cole, ed., *Syntax and Semantics 9: Pragmatics*, 315–332. New York: Academic Press.

Department of Linguistics and Philosophy
32-D808
Massachusetts Institute of Technology
Cambridge, MA 02139

heim@mit.edu