November 2017

When errors aren't: How comprehenders selectively violate Binding Theory

Shayne Sloggett
University of Massachusetts Amherst

WHEN ERRORS AREN'T:
HOW COMPREHENDERS SELECTIVELY VIOLATE BINDING THEORY

A Dissertation Presented

by

SHAYNE SLOGGETT

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2017

Linguistics

WHEN ERRORS AREN'T:
HOW COMPREHENDERS SELECTIVELY VIOLATE BINDING THEORY

A Dissertation Presented

by

SHAYNE SLOGGETT

Approved as to style and content by:

Brian Dillon, Chair

Lyn Frazier, Member

Kyle Johnson, Member

Adrian Staub, Member

Seth Cable, Department Chair
Linguistics

The best lack all conviction, while the worst

Are full of passionate intensity.

–W.B.Yeats

ACKNOWLEDGMENTS

I've been putting off writing these acknowledgments because doing so feels like saying goodbye, and I'm very much not ready to do that. This place has been home for six years now, and being a bit of a homebody, I've grown rather attached to it, and to the people who have made it so. But, with the time drawing close I suppose it's got to be done now, if ever.

So, here we are. First thanks, I suppose, must go to my committee, without whose patient (and only occasionally exasperated) guidance I would not now be writing this. Thank you to Adrian Staub, whose close eye and careful questions have helped push my analysis beyond casual conjecture, and to Kyle Johnson, for years of leading by sartorial example and for putting up with a psycholinguist with delusions of syntactic grandeur. Kyle, our conversations have been alternately vexing and epiphanal; they have made me think harder than I wanted to and I couldn't be more grateful. Likewise, words cannot express the degree of my indenture to Lyn Frazier. Lyn, I can only hope to be half the linguist you have shown me how to be. And finally, Brian Dillon. Brian, you perhaps more than anyone else have shaped the linguist I am today. You were the first person to teach me experimental design, and have have been my constant guide for (somehow) seven years now. I quite literally don't know what it means to work without you. This parting cuts deep.

My committee is not alone, however, in getting me here. I am immensely grateful to John Kingston, for taking me in even when I'm not sure he exists, to Caroline Andrews for her unending patience for my foibles, and to Jeremy Pasquereau for enduring in the face of an endlessly vituperative classmate. I would also like to thank Chuck Clifton, for believing that linguists can always come up with a story, and to Isabelle Charnavel for helping me craft mine. In addition, I wouldn't have made it to graduate school were it not for my phenomenal undergraduate mentors, Donka Farkas, Jorge Hankamer, Jim McClosky, and, especially, Matt Wagers, who convinced me to pursue psycholinguistics in the first place. The department at Santa Cruz was my first exposure to linguistics, and continues to shape the way I see the field. I owe a great deal as well to Sol Lago, Wing Yee Chow, and Dave Kush, who took me in when I was still finding myself as a psycholinguist, and whose continued friendship has been a source of comfort and stability throughout this process. Likewise, Anthony Yacovone and Stephanie Rich, thank you for letting me peeve at you

about the field as you've come into it on your own. You may have gotten more (or less, depending on your perspective) than you bargained for, but your companionship has meant worlds to me.

Outside linguistics, I have been supported by family and friends whose contributions cannot be overvalued. To Florian Gargaillo, I am endlessly grateful for years of grad school commiseration and a shared delight in the absurd. Elaine Teng, we have been friends for over a decade. I would never have made it here without you. Thank you for continuing to see the best in that lonely boy you sat next to in sophomore English. He isn't so lonely anymore. As for family, Maegan, Daphne, and Steve Sloggett, I would not have got here here without the three of you. You were home before UMass was.

And lastly, thank you Amanda Rysling, my friend at camp and the deux of my folie these last five years. I think that about says it all.

ABSTRACT

WHEN ERRORS AREN'T:
HOW COMPREHENDERS SELECTIVELY VIOLATE BINDING THEORY

SEPTEMBER 2017

SHAYNE SLOGGETT

B.A., UNIVERSITY OF CALIFORNIA SANTA CRUZ

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Brian Dillon

It has been claimed that comprehenders use the Binding Theory (Chomsky, 1986) to restrict the search for a reflexive's antecedent in early stages of comprehension (Dillon, Mishler, Sloggett, & Phillips, 2013; Nicol & Swinney, 1989; Sturt, 2003a, 2003b). However, recent findings challenge this view, demonstrating that comprehenders occasionally access antecedents on the basis of their match with a reflexive's morphosyntactic features (Chen, Jäger, & Vasishth, 2012; Patil, Vasishth, & Lewis, 2016; Parker & Phillips, 2017; Sturt, 2003b). In this dissertation, I investigate the source of this 'grammatical fallibility' in the real-time application of Principle A of the Binding Theory. Specifically, I ask whether this pattern of behavior is the direct consequence of an error-prone retrieval mechanism, or if it is instead the result of a discourse-oriented, logophoric interpretation of reflexive forms. This work presents four experiments demonstrating that comprehenders only consider non-Principle A antecedents which act as prominent perspective holders in the discourse. I explain these findings by appealing to local, logophoric center available for reflexive reference.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

REFLEXIVE PRONOUNS IN THEORY AND PROCESSING

In this chapter, I provide an overview of theoretical and experimental investigations of reflexive pronouns, with an emphasis on understanding (*i*) what constraints hold of reflexive interpretations (*ii*) how these constraints might be deployed during real-time sentence comprehension. The chapter is organized as follows. Section 1.1 outlines the standard approach to Binding Theory, focusing on the version advanced in Chomsky (1986). From there, Section 1.2 presents two alternative models to this theory which invoke reflexive predicate formation. In Section 1.2.1, I give a brief characterization of the model proposed by Pollard and Sag (1992), and in Section 1.2.2 an overview of the model presented in Reinhart and Reuland (1993). Section 1.3 then shifts perspectives to examine the real-time application of binding constraints. This section includes a summary of the primary evidence demonstrating strict adherence to Binding Theory, even in early stages of reflexive comprehension, as well as an overview of more recent evidence demonstrating situations in which Binding Theory fails to hold (Parker & Phillips, 2017; Dillon et al., 2013; Nicol & Swinney, 1989; Xiang, Dillon, & Phillips, 2009; Jäger, Engelmann, & Vasishth, 2017). Section 1.3.3 gives a characterization of these findings in terms of a cue-based retrieval memory architecture (Lewis, Vasishth, & Van Dyke, 2006; Lewis & Vasishth, 2005; McElree, 2006), and questions whether this is the most appropriate characterization. Section 1.4 presents a possible alternative interpretation which appeals to logophoric constraints on reflexive interpretation. This section includes a summary of the arguments and data found in Charnavel and Sportiche (2016), as well as an introduction to logophoric pronouns and the interpretations associated with them. Finally, the chapter concludes with a summary of the primary argument of the dissertation, and an overview of the material contained in subsequent chapters.

1.1 Standard Binding Theory

In accounting for reflexive anaphors, many syntactic theories adopt some version of the Standard Binding Theory (SBT; Chomsky, 1986), a simplified version of which is given in (3). In this, and future definitions, "binding" indicates that a referring device (a reflexive or pronoun) is c-

commanded by, and co-indexed, with another node in the structure. Conversely, "free" indicates that a given referring device is *not* bound (i.e. not co-indexed with a c-commanding referent).

(1) **Binding:** A node α binds another node β if α c-commands β, and α and β are co-indexed

(2) **C-command:** A node α c-commands β if the node which immediately dominates α also dominates β

(3) **Standard Binding Theory**

   a. An anaphor (reflexive, reciprocal, or trace) must be bound in its local domain

   b. A pronominal (pronoun or *pro*) must be free in its local domain

   c. An r-expression must be free

At its foundation, this theory expresses the intuition that the referential potential of reflexives and pronouns appears to be in complementary distribution. As seen in (4), an embedded reflexive must refer within the embedded clause, while a pronoun in the corresponding must refer outside it[1]. As a first approximation, then the clause appears to the the "local domain" in which an anaphor must be bound, and a pronominal free.

(4) a. Michael$_i$ noticed that Gob$_j$ had embarrassed himself$_{*i/j/*k}$ yet again.

   b. Michael$_i$ noticed that Gob$_j$ had embarrassed him$_{i/*j/k}$ yet again.

Unfortunately, this definition of locality quickly breaks down in two ways. First, the local domain seems to be smaller than the minimal clause, in some cases. In (5) we see that the local domain can be a DP. The anaphor must refer within the DP, while the pronoun must refer outside it[2], indicating the need for a refined definition of the local domain.

(5) a. Carl$_i$ agreed to send [Tobias'$_j$ pictures of himself$_{*i/j}$] to the talent agency.

   b. Carl$_i$ agreed to send [Tobias'$_j$ pictures of him$_{i/*j}$] to the talent agency.

Second, and perhaps more troublingly, there are situations in which the complementarity between reflexive and pronominal reference appears to break down, as seen in (6) and (7). These examples represent a true break-down in the central intuition of SBT (namely, the complementarity

[1]Note that in this example, the index *k* indicates referents to an utterance-external referent.

[2]I report here judgments as they are reported in the literature. My own intuition is that *himself* in examples like (5a) can refer outside the DP with ease, while the pronominal case in (5b) sounds border-line ungrammatical. However, I may also be broken. Given that reference outside the putative local domain will concern us a great deal later, I set this disagreement in judgmets aside for now in service of presenting the standard theoretical approach to anaphora.

of anaphors and pronominals), according to which either the reciprocals in (6) are referring outside the local domain, or else the pronouns in (7) are referring with it.

(6) a. The children$_i$ heard stories about each other$_i$.

b. The children$_i$ like each other's$_i$ friends.

(7) a. The children$_i$ heard stories about them$_i$.

b. The children$_i$ like their$_i$ friends.

In the following section, I review Chomsky (1986)'s attempt to resolve this conflict within the SBT model. In brief, his solution will be to relativize the definition of locality domain such that it can be smaller for pronominals than for anaphors.

1.1.1 Binding Theory in Chomsky (1986)

Chomsky's amendment to the standard theory is meant to resolve three puzzles for the more basic, standard theory. First, locality must be defined such that possessed NPs act as the local domain for anaphors and pronominals. Second, this definition of locality needs to accommodate the judgments in (8) and (9), wherein the critical anaphor/pronominal is the subject of an embedded clause, and their referential capabilities seem to flip as a function of whether that clause is tensed, or non-finite. Finally, this theory must account for the overlapping distribution of anaphors and pronominals in some DP contexts. In what follows, I present the core components of Chomsky's account interspersed with examples demonstrating the implications of each addition to the theory.

(8) a. They$_i$ would prefer for [each other$_i$ to win].

b. $*$ They$_i$ expect that [each other$_i$ will win].

(9) a. $*$ They$_i$ would prefer for [them$_i$ to win].

b. They$_i$ expect that [they$_i$ will win].

There are four principle components to the system Chomsky proposes: government, the minimal governing category (MGC), binding theory compatibility (BT-compatibility), and a licensing condition on co-indexation. I will address each of the components in turn. The notion of government is the same one taken from Government and Binding Theory (Chomsky, 1981), reproduced here in (10). In brief, government expresses a mutual c-command relation between a category head and a maximal projection, with some stipulations about which categories can act as governors. For our purposes, it will be sufficient to discuss government in terms of containment: i.e.

a governing category is a minimal maximal projection which contains a referring device, and a lexical governor for that device.

(10) **Government:** A category α governs a projection X″ if α and X″ c-command each other

a. If α governs X″, then α governs the specifier and head of X″

b. Subjects and predicates govern one another

c. Only lexical categories and the projections can be governor

d. Infl governs it subject

However, we can already see that this notion of governing category will be too weak, at least as a proxy for the locality domain referred to SBT. For example, given this definition, the governing category of the reflexive in (11) should be the DP: the anaphor *themselves* is c-commanded by the lexical governor *pictures*, making the DP the governing category for the anaphor. If this corresponded to the locality domain for reflexives, SBT would predict this sentence to be ill-formed.

(11) The girls like those [pictures of themselves].

In light of this, Chomsky proposes a slightly different definition of locality which imposes restriction above those imposed by government alone. This notion of "minimal governing category" is given in (12). For the present, we can set aside the clause which stipulates a *BT-compatible* indexation, and focus on the other component of the MGC: the stipulation that the governing category must include a *subject*. Subjects, here, include the specifiers of IP, and DP (i.e. possessors).

(12) **Minimal Governing Category:** A maximal projection containing a subject, and a lexical governor for a referring expression α in which α can be assigned a *BT-Compatible* indexation

With this contingent property of the locality domain in hand, the notion of MGC can readily account for the fact that DPs can act as the binding domain for anaphors and pronominals when possessed, but not otherwise. In other words, we can explain why the DP in (11) is not the binding domain, but the DPs in (5), reproduced in (13), is.

(13) a. Carl$_i$ agreed to send [Tobias'$_j$ pictures of himself$_{*i/j}$] to the talent agency.

b. Carl$_i$ agreed to send [Tobias'$_j$ pictures of him$_{i/*j}$] to the talent agency.

This extension also neatly accounts for the distribution of referring devices in the subject position of embedded tensed and non-finite clauses (examples (8) and (9)). To see how, recall that Infl acts as a governor for its subject. As a consequence, any tensed clause with a subject will constitute a locality domain for binding, meaning that a subject pronominal in a tensed clause

is necessarily free within its domain, while a subject anaphor in a tensed clause is disallowed, because it cannot be locally bound.

That said, appealing to subjects in defining locality domains[3] doesn't explain cases of overlapping distribution for anaphors and pronominals. Chomsky deals with at least one of these cases by suggesting that a hidden PRO element acts as the silent possessor of some DPs, and that the indexation of PRO in these cases determines the choice of referring device. This account is based on the data in (6) and (7), reproduced in (14). Intuitively, the stories told about the children in (14a) are their own stories. In contrast, the stories told in (14b) are the stories of other people. Thus, there is at least some justification for assuming a silent PRO indexed with the appropriate referents (i.e. the children, or some sentence-external referent) sitting in the specifier of the DPs in these examples. Given this indexation, the DP is again the MGC for the anaphor/pronoun, and SBT is satisfied. In (14a), the anaphor *each other* is bound within its MGC with the possessor PRO, while in (14b) the pronoun is free within its MGC.

(14) a. The children$_i$ heard [PRO$_i$ stories about each other$_i$].

b. The children$_i$ heard [PRO$_j$ stories about them$_i$].

This leaves one final outstanding datum to be accounted for: the overlapping distribution of possessive pronominals and anaphors, reproduced in (15). Working within the framework of keeping anaphors locally bound and pronominals localy free, we intuitively want the entire sentence to be the locality domain of the anaphor in (15a), but DP to be the locality domain of the pronominal in (15b). If this could be accomplished, then the fundamental complementary distribution of anaphors and pronominals could be maintained: anaphors are always locally bound, and pronominals are always locally free. Overlapping distribution, then, arises because sometimes the locality domain for pronominals is smaller than that for reflexives.

(15) a. The children$_i$ like each other's$_i$ friends.

b. The children$_i$ like their$_i$ friends.

Chomsky accomplishes this by incorporating *BT-compatibility* into the definition of the minimal governing category. The definition of BT-COMPATIBILITY is given in (16), and, descriptively, is essentially identical to the definition of binding theory in SBT. Given some definition of locality, a particular indexation scheme is compatible with the Binding Theory if (a) an anaphor is locally bound, (b) a pronoun is locally free, or (c) an r-expression is globally free. However, by including

[3]Chomsky justifies this stipulation by appealing to the Specified Subject Condition (SSC).

this notion in the definition of the MGC, Chomsky effectively relativizes the possible size of the locality domain for a given referring device *to that referring device*. Thus, the locality domain for a reflexive is now the smallest maximal projection containing: a subject, a governor, and a *possible indexation in which the reflexive is locally bound*. Likewise, the locality domain for a a pronoun is the smallest maximal projection containing a subject, a governor, and a possible indexation in which the pronoun is locally free.

(16) **BT-compatibility:** an indexation I is *BT-compatible* with a referring expression α and a locality domain β if:

a. α is an anaphor and is bound in β under I

b. α is a pronominal and is free in β under I

c. α is an r-expression and is free in β under I

Returning to the problem posed by (15), we can see that this relativization of the size of the locality domain handily produces a smaller domain for pronouns than reciprocals. In (15b), the DP contains a subject (the possessive pronoun itself), a governor for the pronoun (the head noun[4]), and a BT-COMPATIBLE indexing: the pronoun can be assigned any index, as no referent in that domain can locally bind it (i.e. any indexation in this domain renders the pronoun locally free). Thus, the DP acts as the MGC for the possessive pronominal. In contrast, the reciprocal in (15b) requires a larger locality domain. In this case, though the DP contains a governor for the reciprocal (again, the head N), and a subject (the reciprocal itself), there is no indexation within this domain which would satisfy the *BT-compatibility*: as, just as with the pronoun *there is no referent which could possibly locally bind it*. Thus the MGC must be extended until it includes a potential local-binder for the reciprocal, with the result that the entire sentence acts the locality domain.

Lastly, we get to the indexation licensing condition, the final component of this binding model. The purpose of this mechanism is simply to act as a means of enforcing the previous three constraint descriptions. Notably, government and *BT-compatibility* are descriptive relations used in defining the MGC. The MGC itself does not enforce Binding Theory, rather, it defines the domain within which anaphors/pronominals should be bound or free respectively. The indexation licens-

[4]Given that the head noun doesn't c-command the possessor position, at least in its surface position in more modern DP syntax, I'm assuming that this is by analogy to the fact that Infl governs its subject. Alternatively, one could suppose that clausal subjects are base-generated below I, and that possessors are generated beneath D, and that it is these heads which are responsible for governing the traces of the subject/possessor phrases. On the whole, this may just be a point of incompatibility between somewhat out-dated notion of government, and more current syntactic theories. This will be addressed, to some extent, in our later consideration of modern approaches to Binding Theory (Charnavel & Sportiche, 2016)

ing condition thus insists that the locality domain of a given indexation scheme must correspond to the MGC of the referring device (in the case of anaphors and pronominals[5]).

(17) **Indexation Licensing Condition:** An indexation I is licensed for a referring expression and a locality domain β if either:

a. α is an r-expression *and*:

i. Either: α heads its chain, and β is the utterance

ii. Or: β is the domain of the head of the chain of α

b. α is an anaphor or pronominal, and β is the minimal governing category of α

There is one final case which remains problematic for this theory, however, given here in (18). This sentence is analogous to the finite-clause cases we saw above (example (8), in which a reciprocal could not act as the subject of a tensed embedded clause. There, the explanation was that the tensed embedded clause acted as a locality domain for the reciprocal because it contained a governor (the embedded Infl head), and a subject (the reciprocal). Given the addition of BT-COMPATIBILITY as a condition on defining the MGC, we now have to augment this picture to allow that Infl can also be co-indexed with the anaphor, and that this satisfies the local binding condition of Principle A[6]. Regardless, the point remains: an embedded tensed clause should be the MGC for any anaphor in (or embedded in) the embedded subject position. Why, then , can the anaphor in (18) be bound by the matrix subject?

(18) The children$_i$ thought that [[pictures of each other$_i$] were on sale].

To explain this datum, Chomsky appeals to the "*i*-within-*i* condition", which prevents a phrase phrase from being co-indexed with a phrase which contains it. In light of this constraint, he points out that in order for the embedded clause to satisfy the *BT-compatibility* requirement, *each other* must be co-indexed with the embedded Infl head. However, the phrase containing the anaphor (i.e. the embedded subject) is *itself* already co-indexed with Infl (for agreement reasons). Given this, co-indexing both phrases with Infl would result in a violation of the "*i*-within-*i* condition", as *each other* would bear the same index as the phrase containing it (the embedded subject). There-

[5]The licensing condition does slightly more for r-expressions, which are not of concern to us here.

[6]Blame Chomsky, not me.

fore, indexing *each other* with the embedded Infl is, in this case, rule out on independent grounds, and the MGC must be extended to include the matrix subject (the intuitive binder)[7].

As will be noted shortly, this solution has not been satisfying for many others in the literature. In particular, it struggles to accommodate the apparent acceptability of a related set of sentences, of which (19) is a representative example. In this example, the critical anaphor (here, the reflexive *themselves*) is *not* contained in the embedded subject. Thus, co-indexation with Infl should not induce an *i*-within-*i* violation. Moreover, there's a perfectly good (albeit, morphosyntactically illicit) binder in the embedded subject. In either event, the embedded clause in this case *should* constitute the MGC for the anaphor, and so to the extent that this sentences is grammatical, it poses a problem a problem for Chomsky's system. This, and other difficulties shall be explored in more detail in the following section.

(19) The children thought that the newspaper had printed pictures of themselves last week.

1.1.2 Standard Binding Theory: a summary

In this section I have attempted to lay out the central insights and data associated with the Standard Binding Theory approach. At its core, this theory takes the complementary distribution of anaphors and pronominals to be the primary data in need of an explanation. As such, it explains the distribution of reflexives and pronouns as a function of the need of the former to be locally bound, and the latter to be locally free. As we saw, this basic picture captures many of the basic facts about anaphors and pronominals, but it also faces non trivial challenges in case where (1) there is variability in the size of the locality domain for a given referring device, or (2) there is overlap in the distribution of referring devices. I then presented the system proposed by Chomsky (1986) to demonstrate one of the more influential ways in which these challenges have been approached. In brief, problem (1) was addressed by including the requirement that locality domains include a subject. Problem (2) was solved by relativizing the size of the locality domain associated with a referring device to that particular referring device. Overall, these solutions gain considerable traction, and allow us to preserve the primary insight of complementary

[7]Chomsky considers a variation on this approach in which anaphors are always LF-moved (in English) to adjoin to Infl (or V, in the case of object-binding). This allows him to remove the stipulation that Infl can act as the antecedent, thereby making embedded tensed clauses too small as the MGC for anaphors contained in embedded subjects, and deriving the larger MGC for examples like (18). To explain the fact that anaphors cannot *themselves* act as embedded subjects, he notes that at LF, the anaphor will have been moved from this position, leaving in its place a trace, and resulting in a violation of the ECP principle. While this solution removes many of the stipulations found in the version described in more detail here, it runs into the same problems as the earlier account, which I enumerate shortly.

distribution between anaphors and pronominals. However, this system is not without its flaws and alternatives, both of which are explored in the next section.

1.2 Predicate-based theories of binding

One long-standing competitor for the SBT approach has been so-called "predicate based" theories of binding. Rather than starting from the observation of complementary distribution, these models suggest that local-binding arises from the creation of "reflexive predicates"—special syntactic/semantic objects which require particular morphology to license them. Consequently, these theories are not concerned with defining locality domains, or (directly) deriving complementary distribution among referring devices. Instead, they focus on describing constraints on reflexive predicates as a means of predicting when non-local binding should be possible. Differences among referring devices falls out as a function of the predicates a given device's morphology allows it to participate in.

In this section, I present two different analyses in this tradition. Both models point to similar failings of the SBT model, and both address these failings by appealing to a predicate-based explanation of binding, rather than one based on locality restrictions. However, their implementations of this intuition remain fairly distinct, and so worth considering individually. First, I present the model proposed by Pollard and Sag (1992), who propose reflexive pronouns must find their antecedents among the co-arguments of their syntactic predicates (if such referents exist). This model is couched in terms of Head-driven Phrase Structure Grammar (HPSG), and so represents a fairly marked departure from the SBT model in more than one respect. The second model comes from Reinhart and Reuland (1993), who suggest that the formation and interpretation of reflexive predicates is the fundamental role of binding theory.

1.2.1 Binding with co-arguments: Pollard and Sag (1992)

Pollard and Sag (1992) present a compelling critique of the SBT model, noting several discrepancies between the theories predictions, and observed behavior with reflexive and reciprocal anaphors. In particular, Pollard and Sag (1992) identify five key predictions of SBT:

(*i*) Anaphors in English should be bound sentence internally

(*ii*) Anaphors in English cannot be discourse bound (sic. (*i*))

(*iii*) Anaphors in English cannot have split antecedents

(*iv*) Binding theory makes no distinction between *themselves* and *the picture of themselves* (etc.)

Relying primarily on the evidence in (20)-(22), they show that each of these predictions is disconfirmed. I present each prediction and corresponding piece of counter-evidence below.

(20) a. The picture of himself$_i$ in the museum bothered John$_i$.

b. The picture of herself$_i$ on the front page of the *times* made Mary's claims seem somewhat ridiculous.

c. John$_i$'s intentionally misleading testimony was sufficient ot ensure that there would be pictures of himself$_i$ all over the morning papers.

(21) a. John was going to get even with Mary. That picture of himself in the paper would really annoy her, as would the other stunts he had planned.

b. "Whom he$_i$ [Phillip] was supposed to be fooling, he$_i$ couldn't imagine. Not the twins, surely, because Désirée, in the terrifying way of progressive American parents, believed in treating children like adults, and had undoubtedly explained to them the precise nature of her relationship with himself$_i$." (David Lodge)

(22) a. The agreement that [Iran and Iraq]$_i$ reached guaranteed each other$_i$'s trading rights in the disputed waters until the year 2010.

b. John$_i$ asked Mary$_i$ to send reminders about the meeting to everyone on the distribution list except themselves$_i$.

Examples (20)-(22) directly refute predictions(*i*)-(*iii*), as laid out by Pollard and Sag. In (20), we see that at least some instances of anaphors need not be internally bound: each of the reflexives in these sentences cannot be bound in the sense of Chomsky (1986). In (20a) and (20b), the critical reflexive precedes its antecedent, which is itself the complement of the matrix predicate. It might be possible to suggest that, in these cases, the offending anaphor is generated low in the structure (critically, below its antecedent), and then moved to this higher position, but example (20c) puts the lie to this hope, as the reflexive in this example is co-referent with the possessor of the main-clause subject. Thus, without a significant restructuring of the notion of binding, these examples seem fundamentally at odds with SBT.

More strikingly still, the examples in (21) demonstrate that the antecedent of a reflexive need not even be sentence-internal. These examples demonstrate instances of what Pollard and Sag term "discourse binding"—situations in which the reflexive finds its antecedent sentence-externally from the local discourse context. While some of these examples (e.g. 21b) are of a

decidedly literary flavor, they nevertheless demonstrate the potential of reflexive forms to remain internally un-bound, in some situations, and provide further apparent counter-evidence to SBT.

Finally, the examples in (22) show cases of split-antecedents co-determining the antecedent of an anaphor. In the case of (22a), the argument might be made that it is the conjoined DP, rather than each referent individual, which binds the anaphor[8], but this argument is not possible for (22b), in which the intended antecedents are the subject and object of the matrix predicate. Thus, (22b), at the very least, constitutes true evidence of split-antecedence for English anaphors (in addition to demonstrating non-local binding). That split-antecedence should be ruled out on SBT is, at first, not transparently obvious. Howver, Pollard and Sag point out that split antecedence would entail more than a single constituent binding an anaphor, thus contradicting the statement that an anaphor must have *a* binder within its locality domain. However, this seems to impose a rather strict reading of the requirement that an anaphor be locally bound, one which doesn't seem present in, at least, the formulation of Chomsky (1986)'s theory laid out in the previous section. At worst, this system does not *predict* the possibility of split antecedence, though it doesn't strictly rule it out. Thus, we might consider examples like (22) evidence that SBT needs to be augmented in some fashion to account for split-antecedent interpretations, but not as counter evidence against its core claims.

Regardless, with these examples Pollard and Sag present strong evidence against one of the core intuitions expressed in SBT: that anaphors need to be locally bound. In the most extreme cases, it seems, may even be bound cross-sententially, an observation which is fundamentally incompatible with the SBT model. This leads to Pollard and Sag's final contention with SBT: that bare anaphors (e.g. *themselves* in direct object position) should be treated identically to anaphors embedded in other constituents (e.g. *themselves* in *the picture of themselves*). In the previous section, we saw that SBT deals with differences among anaphors in these environments by stipulating the need for a "subject" to establish a locality domain, and by relativizing the size of the locality domain to the particular referring device. Thus, according to SBT, there is no meaningful difference between *themselves* and *picture of themselves*: in both cases the anaphor is constrained by a fundamental need to be locally bound. In contrast, Pollard and Sag contend that this formalization misses the fundamental generalization about such cases—that only bare argument anaphors *require* local binding, while reference is much freer for non-argument anaphors. This argument is based on contrasts like those in (23) and (24).

[8]Though note that conjoined DP is in a structurally inappropriate position to bind the anaphor

(23) a. ∗ Gob said that the newspaper article deeply embarrassed himself.

b. Gob said that the newspaper article included deeply embarrassing pictures of himself.

(24) a. ∗ Lucille complained to Michael that the puff piece had unexpectedly mentioned themselves.

b. Lucille complained to Michael that the puff piece had unexpectedly mentioned the evidence against themselves.

Given these examples, Pollard and Sag note that when local binding is *required*, the anaphor is "in the same syntactic argument structure as its binder". Put differently, an anaphor is only obligatorily locally bound when it is coargument with some other referent[9]. On the basis of this observation, they suggest a re-configuration of Binding Theory, given informally here in (25), where argument-obliqueness is understood in terms of the hierarchy in (26).

(25) **Principle A:** An anaphor must be coindexed with a less oblique co-argument, if one exists

(26) **Argument Obliqueness:** SUBJECT≪PRIMARY OBJECT≪SECOND OBJECT≪OTHER COMPLEMENTS

With this (simplified) definition in hand, we have a reasonably straightforward means of capturing the SBT inconsistent data in (20)-(24). In each instance of locally-unbound anaphora, the anaphor is not part of the same syntactic argument structure as some less oblique argument. Instead, it is embedded in a constituent which is, itself, an argument of some larger predicate. Pollard and Sag term anaphors in these environments EXEMPT ANAPHORS, given that they appear to be *exempt* from the local binding requirement. Importantly, this definition still allows us to capture the facts for anaphors embedded in possessed picture NPs—as long as possessors still act as subjects (a fact independently needed in SBT), they will count as co-arguments for an embedded anaphor, and thus necessarily bind it. In what follows, I briefly recapitulate the manner in which Pollard and Sag manipulate "coargumenthood" to achieve this result, before turning to an examination of non-binding constraints on exempt anaphors.

1.2.1.1 Defining coargumenthood

For Pollard and Sag, coargumenthood is defined in terms of the arguments subcategorized by predicates: if two elements act as members of the same categorization frame, they belong to the same predicate and are coarguments of each other. Furthermore, in HPSG (Pollard and Sag's

[9]Where the notion of "coargument" is yet to be nicely (in the sense of Agnes Nutter) defined.

preferred idiom), subcategorization plays an important role both in building syntactic structure, and in establishing arguments' relative obliqueness. Given this, I will briefly summarize subcategorization in HPSG before presenting Pollard and Sag's revized version of the binding theory.

In HPSG, lexical items introduce "subcategorization frames", ordered lists specifying the complements necessary for constructing a grammatically complete syntactic projection. The notion of "complement", here, is fairly broadly construed, and includes the subjects of predicates and possessors of NPs. While subcategorization frames are ordered, this order does not correspond to the surface positions of the argument involved, but rather these arguments' relative obliqueness, following the cline given in (26). Importantly, lexical items in HPSG subcategorize both the categories of their complements, as well as their "content". Content, broadly speaking, contains thematic role information and index assignment. Thus, for example, the (slightly simplified) lexical entry for a verb like *chased* is given in (27), where *chased* subcategorizes for two NP arguments corresponding to the agent and patient thematic roles specified in the verb's content.

(27)
$$\begin{bmatrix} \textsc{category} & \begin{bmatrix} \textsc{head} & \text{V} \\ \textsc{subcat} & < \text{NP}{:}x, \text{NP}{:}y > \end{bmatrix} \\ \textsc{content} & \begin{bmatrix} \textsc{relation} & \textit{chase} \\ \textsc{agent} & x \\ \textsc{patient} & y \end{bmatrix} \end{bmatrix}$$

Finally, the Subcategorization Principle requires that heads combine with complements in such a way that each complement corresponds to one member of the head's subcategorization list. An example of this principle in action is given in Figure 1.1 for the verb *chased*. Note that nodes in this structure can be identified with the nodes of more traditional trees (e.g. S, VP, NP, etc), suggesting a relatively straight forward mapping between this representation, and those used in the previous section (modulo the formulation of subcategorization/projection).

Using this notion of subcategorization, Pollard and Sag have a fairly straightforward way of capturing the intuition expressed in (25). If an anaphor is a member of a subcategorization list, it must refer to a higher member on that list *if one exists*. If the anaphor is the first, or only, member of the subcategorization list, then it need not be bound. They formalize this constraint using the set of principles in (28)-(30). Here, the notion of o-command replaces c-command from SBT, but is defined in terms of subcategorization lists. O-binding, then, is simply our previous definition of binding substituting o-command in place of c-command. The critical caveat allowing

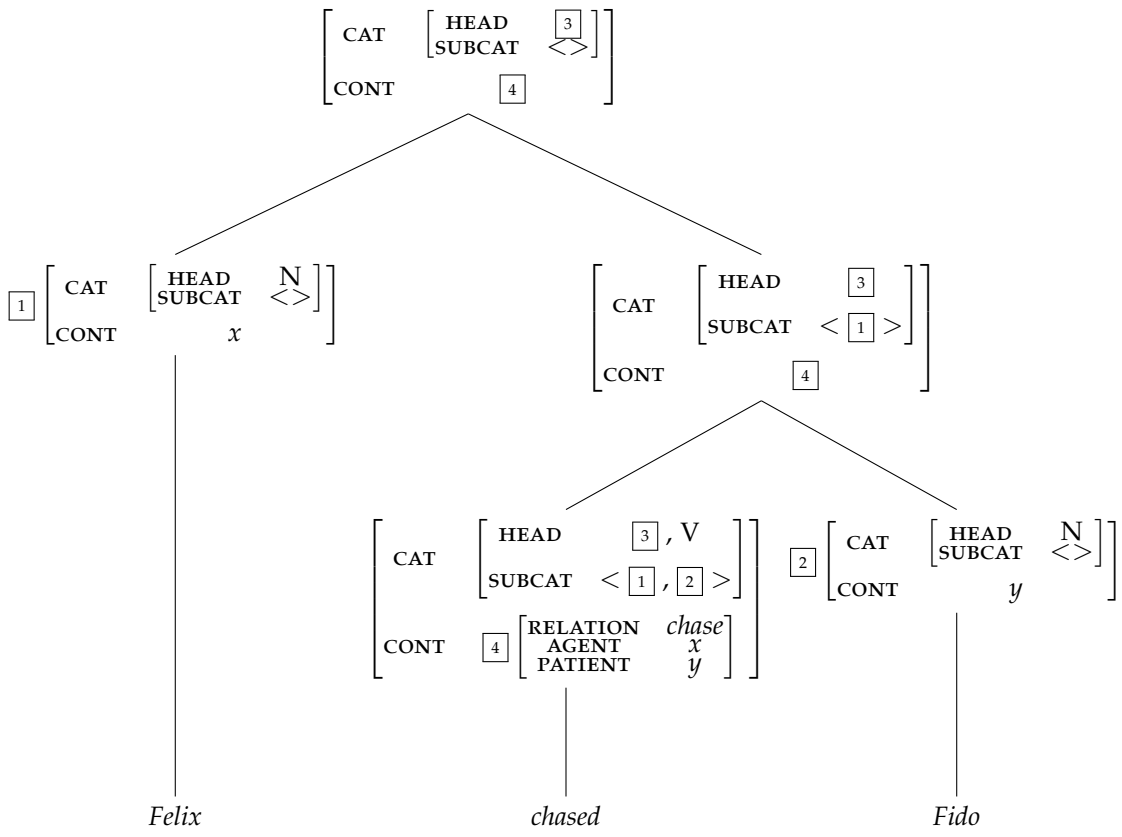


Figure 1.1. A sample tree-structure in HPSG, demonstrating subcategorization and argument saturation based on the Subcategorization Principle

non-local reference is introduced with the new definition of Principle A, as only an anaphor which is o-commanded must be o-bound[10].

(28) **O-command:** α o-commands β if the content of α is a referential parameter, and there is a SUBCAT list on which α precedes β

(29) **O-binding:** α o-binds β if α o-commands β and α and β share an index

(30) **Principle A:** An o-commanded anaphor must be o-bound

This formalism captures many of the same generalizations noticed in our discussion of SBT. Notably, it correctly predicts that: direct object reflexives will necessarily be bound by their subject (31); prepositional objects may be bound by either the direct object, or the subject (32); and that

[10]Additional machinery is introduced to accommodate binding into prepositional phrases (e.g. *Brian talks to himself*, wherein the content of the prepositional phrase must be identified with the content of the NP it contains. This ensures that the verb's subcategorization frame includes the referent of its subject and the referent of its prepositional object, thereby allowing the binding principle in (30) to be enforced.

more oblique prepositional objects may be bound by less oblique arguments, but not vice versa (33).

(31) [SUBCAT $<$ NP:$npro_i$, NP:ana_i $>$]

a. John hates himself.

b. $*$ John thinks Mary hates himself.

(32) [SUBCAT $<$ NP_i , NP_j , PP:$ana_{i/j}$ $>$]

a. Mary explained Bill to himself.

b. Mary explained Bill to herself.

c. $*$ John forgot that Mary had explained Susan to himself.

(33) [SUBCAT $<$ NP_i , $PP[to]_j$, PP[about]:$ana_{i/j}$ $>$]

a. Mary talked to John about himself.

b. Mary talked to John about herself.

c. $*$ Mary talked about John to himself.

d. $*$ Mary talked to himself about Bill.

One key feature of this system is that it derives the fact that anaphors in some positions can be locally free. Unlike SBT, it does not require that *all* anaphors be locally bound, only those which are locally o-commanded. Thus, the theory correctly predicts that anaphors in select syntactic positions will be referentially free (i.e. "exempt"), as observed in the examples which began this section. This will occur whenever the anaphor is: (*i*) the only referential item on its subcategorization list; (*ii*) the *highest* item on its subcategorization list. Examples (34) and (35) give the anaphor-containing subcategorization frames for sentences in which the anaphor is exempt.

(34) [SUBCAT $<$ DetP , PP[of]:*ana* $>$]

a. The picture of himself in the museum bothered John.

b. John's intentionally misleading testimony was enough to ensure that there would be pictures of himself all over the morning papers.

c. John was going to get even with Mary. That picture of himself in the paper would really annoy her, as would the other stunts he had planned.

(35) [SUBCAT $<$ NP[*poss*]:*ana* $>$]

The agreement that Iran and Iraq reached guaranteed each other's trading rights in the disputed waters until the year 2010.

Note that this still captures the fact that anaphors in possessed NPs will *not* be exempt. As in SBT, the possessor acts as a subject for the NP, and therefore will co-habit the anaphor's subcategorization list in a less-oblique position (and consequently o-command it). Example (36) shows this technically.

(36) [SUBCAT $<$ NP[*poss*]:*npro* , PP[of]:ana_i $>$]

Jim$_i$ appreciated John$_j$'s picture of himself$_{*i/j}$.

With respect to how exempt anaphors actually find their antecedents (accidental co-indexation, discourse-guided search, etc.), Pollard and Sag remain relatively agnostic. However, they do note several descriptive constraints which seem to hold of exempt anaphoric reference. We will return to these constraints shortly.

Another important difference between this theory and the SBT model arises from the manner in which they treat the contrast between finite and non-finite embedded clauses, repeated here in (37). Recall that in SBT, this contrast fell out from the fact that a tensed clause, but not a tense-less one, will constitute a locality domain for an subject anpahor. Thus, in (37b), the anaphor remains locally unbound and therefore ungrammatical[11]. Importantly, on this account, the anaphor in (37a) *is* locally bound by the matrix subject—the locality domain in this case is simply the sentence.

(37) a. They would prefer for each other to win.

b. $*$ They expect that each other will win.

Pollard and Sag's system assigns a different explanation to this contrast. First, under their theory, the anaphor in (37a) is critically *free*. As the subject of the embedded clause, it is he least oblique referent of the embedded verb's subcategorization frame, and thus free to remain unbound. Consequently, reference to the matrix subject is the product of "accidental" co-indexation, rather than o-binding, directly. They argue that this view of subject anaphors in embedded non-finite contexts is the correct one based on examples like (38), in which these anaphors refer to referents in illicit structural positions (according to SBT). Given this, they argue that their system correctly predicts that subject-anaphors in non-finite clauses should *not* be locally bound, while SBT wrongly predicts that they *must* be.

(38) a. What John would prefer is for himself to get the job.

b. The thing Kim and Sandy want most is for each other to succeed.

[11] Alternatively, under Chomsky (1986)'s movement account of anaphors, this is a violation of ECP.

This leaves this issue of explaining the ungrammaticality of examples like (37b). Here Pollard and Sag appeal to morphology, noting that English anaphors lack a nominative form. Presumably, a similar argument must be given for why reflexive pronoun cannot occupy the possessor position of an NP, and so this explanation of the contrast in (37) seems somewhat independently supported.

One possible point on which this theory founders lies in potential differences between animate, and inanimate reflexives. Notably, non-local reference is intuitively much less available if the anaphoric device is inanimate, as seen in (39). Assuming that *itself* and *himself* are not a fundamentally different objects, this contrast is puzzling under Pollard and Sag's theory. The anaphors in both cases should be exempt, and yet reference from this position with an animate reflexive is acceptable, while reference with an inanimate is not. A similar observation is made for French anaphors by Charnavel and Sportiche (2016), who note that exempt behavior is never observed with the inanimate anaphors *son propre* and *elle-même*. This is data to which we will return in Section 1.4. For the present, we note (39) as an interesting contrast currently unexplained by the current system[12].

(39) a. John's email suggested that opinions about himself would be divided.

b. $*$ The email's subject line suggested that opinions about itself would be divided.

1.2.1.2 Constraints on exempt anaphors

As noted above, Pollard and Sag do not give an explicit account of how exempt anaphors find their antecedents. However, they do report several constraints which seem to hold of exempt anaphoric reference. I recapitulate these in brief, here. First, Pollard and Sag note an apparent blocking effect for non-local interpretations, reconstructed here in (40) and (41). If two animate referents precede an exempt anaphor, it must be anteceded by the structurally closer referent. However, if one of the referents is inanimate, the anaphor refers to the animate referent, regardless of distance. Two points are worth noting, here. First, Pollard and Sag report report these blocking effects as though they are categorical, though evidence presented later in this work demonstrates that they are not. Second, the authors suggest that processing factors are responsible for this apparent blocking effect, though they offer little speculation about the nature of these factors. Both of these points will occupy us greatly in Chapter 3 of this work, where we will see evidence

[12]In fairness, these data are similarly mysterious under SBT. Recall that Chomsky invoked the i-within-i constraint to explain the acceptability of examples like (39a). Why this explanation should not extend to (39b) is mysterious.

that blocking effects are not categorical, and are the product of perspective-taking mechanisms in sentence comprehension.

(40) a. Bill remembered that Tom saw a picture of himself in the post office.

b. Bill remembered that Tom said that there was a picture of himself in the post office.

(41) a. Bill remembered that the *Times* had printed a picture of himself in the Sunday edition.

b. Bill suspected that the silence meant that a picture of himself would soon be hung on the post office wall.

Relevant to this latter point, Pollard and Sag also note that long distance reference is facilitated if the discourse is reported from the intended antecedent's point of view. That is, exempt anaphors preferentially refer to the "perspective holder" of a discourse. They present (42) as evidence of this claim. Example (43) demonstrates a similar point: assuming that a clause must be interpreted from a *single* perspective, and that exempt anaphors refer to this perspective holder, we expect that multiple exempt anaphors within a clause should necessary refer to the same antecedent. That is, clause-mate exempt anaphors are expected to "shift together" (Anand, 2006). To the extent that (43) is ungrammatical, then, we have evidence that exempt anaphors are tracking the perspective holder of the clause[13].

(42) a. John was going to get even with Mary. That picture of himself in the paper would really annoy her, as would all the other stunts he had planned.

b. $*$ Mary was taken aback by all the attention John was receiving. That picture of himself in the paper had really annoyed her, and there wasn't much she could do about it.

(43) $*$ John traded Mary pictures of herself for pictures of himself.

Finally, Pollard and Sag present evidence of perspective sensitivity in the impact of psych predicates on the referential capabilities of exempt anaphors. Notably, these verbs introduce an *experiencer* thematic role, and the assignee of this role is, intuitively, a salient perspective for interpreting a clause. As seen in (44a), an exempt anaphor is preferentially interpreted as referring to the the experiencer role of a psych-predicate (and thus, the relevant perspective for the clause).

[13]Personally, I do not find this example terribly unacceptable. If this sentence were to be acceptable, however, it might still be possible to align these facts with persective taking in exempt anaphora. Notably, this kind of "shift together" constraint is also reported for non-local interpretations of the Mandarin reflexive *ziji*. However, in this case, the shift-together constraint can be obviated *if* one of the antecedents locally binds the anaphor. Thus, "shift together" is only enforced when both instances of the anaphor find their antecedent non-locally. In this example, we might suppose that either anaphor finds its antecedent via conventional binding means, while the other refers to the perspective holder.

Crucially, this isn't a problem with referring to possessors, as seen in (44b), indicating the salience of the experiencer role for determining reference[14].

(44) a. The picture of himself$_{*i/j}$ in the newspaper bothered John$_i$'s father$_j$.

b. The picture of himself in the newspaper dominated John's thoughts.

Interestingly, this sensitivity to perspective-setting properties in a sentence bear a striking resemblance to the behavior of logophoric pronouns in other languages (Kuno, 1972; Sells, 1987; Culy, 1997; Charnavel & Sportiche, 2016). Likewise, such properties look very like constraints on long-distance interpretations of reflexive pronouns in other languages (e.g. Mandarin, Japanese, and Icelandic; Huang & Liu, 2001; Cole & Wang, 1996, i.a.). This similarity will occupy much of the remaining discussion in this dissertation.

1.2.1.3 Summary

In this section, I have presented the work of Pollard and Sag (1992). Empirically, this work presents serious challenges for the traditional SBT model, demonstrating that not all anaphors in English need to be locally bound. In fact, when the anaphor is the sole, or least oblique, argument of its predicate, it may be locally free. The authors capture this intuition with a formalism rooted in HPSG, suggesting that anaphors are locally bound when they are preceded by other arguments in an HPSG subcategorization frame. As we saw, this formalism correctly predicts those situations in which an anaphor may be "exempt", and find its antecedent non-locally. Moreover, we have seen that while exempt anaphors may be syntactically free (in the sense that they are not subject to Principle A), they are nonetheless subject to discourse constraints on their interpretation. In particular, anaphors in these positions seem to preferentially seek out the perspective holder of an utterance in a manner which closely mimics the behavior of long-distance reflexives and logophoric pronouns cross-linguistically. We turn now to another theory aimed at explaining exempt anaphors.

1.2.2 Reflexive predicates: Reinhart and Reuland (1993)

The basic preoccupation of Reinhart and Reuland (1993) is fundamentally similar to that of Pollard and Sag (1992). In both cases, the authors are concerned with uses of apparently reflexive

[14]This sensitivity to perspective might be used to explain the contrast noted in (39). Inasmuch as inanimate reflexives cannot hold a perspective, we might expect them to be illicit in contexts which are preferentially associated with a perspective center. However, while this is a plausible description of the contrast between animate/inanimate anaphors, it does not itself derive it, and we are still left to wonder how inanimate anaphors in exempt environments might get bound.

anaphors which appear to be locally unbound in contravention of the SBT model. However, where Pollard and Sag's preoccupation is English, and the theory couched in terms of subcategorization frames in HPSG, Reinhart and Reuland's work primarily draws evidence from Dutch, and their theory is focused on the creation of "reflexive predicates". This section, then, is organized as follows: first, I present the primary data from Dutch relevant to understanding Reinhart and Reuland (1993)'s theory; second, I present the core elements of the theory and discuss its various successes and shortcomings; finally, I present a comparison with Pollard and Sag (1992)'s model, as the two theories capture much the same intuition with fairly different technologies.

1.2.2.1 Dutch: The facts were these

There are three referring devices in Dutch which primarily occupy Reinhart and Reuland's thinking: *zichzelf*, *zich*, and *hem*. Relative to English, *zichzelf* behaves roughly like a standard reflexive anaphor: it must (usually) be locally bound. Similarly, *hem* behaves very much like English pronominals, and must be locally free. Problematically the distribution of *zich* overlaps with both of these categories: *zich* usually cannot be locally bound, but sometimes can. Consequently, the Dutch pronominal system poses a challenge for the complementarity assumption underlying the SBT model. How do we deal with a referring device which is, so to speak, neither fish nor fowl? Before jumping to Reinhart and Reuland's solution, I present the primary facts from Dutch, below. Throughout, this presentation *zichzelf* and zich are glossed as SELF and SE, respectively, for reasons which will become clear shortly.

In simple, transitive sentences, *zichzelf* may be used to refer to the subject, but neither *zich*, nor *hem* may be (45). Likewise, for thematic prepositional objects, *zichzelf* must be used for subject reference (46). In contrast, if the prepositional object is predicative (e.g. *behind*, *in front of*), then the distributions flip: *zich* and *hem* may be used for subject reference, but *zichzelf* may not (47). Based on these examples, then, we see that *zich* and *hem* substantially overlap in their distribution, and do not appear to overlap with *zichzelf*.

(45) Jan_i haat zichzelf$_i$/$*$zich$_i$/$*$hem$_i$.

Jan$_i$ hates SELF$_i$/$*$SE$_i$/$*$him$_i$
Jan$_i$ hates himself$_i$/$*$him$_i$.

(46) Max_i praat met zichzelf$_i$/$*$zich$_i$/$*$hem$_i$.

Max$_i$ speaks with SELF$_i$/$*$SE$_i$/$*$him$_i$
Max$_i$ speaks with himself$_i$/$*$him$_i$.

(47) Klaas$_i$ duwde de kar voor hem$_i$/zich$_i$/*zichzelf$_i$ uit.

Klaas$_i$ pushed the cart before him$_i$/SE$_i$/*SELF$_i$ out
*Klaas$_i$ pushed the cart in front of him$_i$/*himself$_i$*

However, this complementary distribution flips when we consider inherently reflexive predicates (e.g. *behave*, *be ashamed*) and ditransitive structures. For inherently reflexive prdicates, *zichzelf* or *zich* must be used, while *hem* may not be (48). Similarly, in ditransitive sentences, if both objects refer to the subject, then one must be *zichzelf*, but the other may be *zich*, however neither may be *hem* (49). Curiously, there is variation in the acceptability of *zichzelf* in either of both of these configurations. For inherently reflexive predicates, langauges like German permit the SE-anaphor *sich*, but not the SELF-anaphor *sichselbst*. Likewise, a ditranistive structure in which both object arguments are realized as *zichzelf* in Dutch is marked, relative to sentences in which one argument is realized as *zich*. Thus, the claim that *zich* and *zichzelf* have truly overlapping distribution is somewhat dubious. Nevertheless, it remains true that *zich* has a wider distribution than *hem*, and may on occasion be locally bound.

(48) Max$_i$ schaamt zichzelf$_i$/zich$_i$/*hem$_i$.

Max$_i$ shames SELF$_i$/SE$_i$/*him$_i$
*Max$_i$ is ashamed (of himself$_i$/*him$_i$).*

(49) a. Henk$_i$ wees zichzelf$_i$ aan zich$_i$/*hem$_i$ toe.

Henk$_i$ assigned SELF$_i$ to SE$_i$/*him$_i$
*Henk$_i$ assigned himself$_i$ to himself$_i$/*him$_i$*

b. Henk$_i$ wees zich$_i$/*hem$_i$ aan zichzelf$_i$ toe.

Henk$_i$ assigned SE$_i$/him$_i$ to SELF$_i$
*Henk$_i$ assigned himself$_i$/*him$_i$ to himself$_i$.*

c. ? Henk$_i$ wees zichzelf$_i$ aan zichzelf$_i$ toe.

Henk$_i$ assigned SELF$_i$ to SELF$_i$

In brief, the data in (45)-(49) show the intermediate status of the SE-anaphor *zich*. While *zichzelf* must be locally bound, and *hem* may not be, *zich* appear to allow local binding in some cases, but not in others. This posese a problem for the SBT model, which does not *a priori* allow for a pronominal of intermediate status, and constitutes the central challenge Reinhart and Reuland aim to address.

1.2.2.2 Binding with predicates

To account for the intermediate status of *zich*, Reinhart and Reuland parameterize referring expressions along two dimensions: a *reflexivizing function*, and *referential independence*. According to this characterization, only *zichzelf* is a reflexivizing function, and only *hem* is referentially independent.

Table 1.1. Parameters for referring devices (Reinhart & Reuland, 1993)

	SELF	SE	PRO
Reflexivizing Function	+	−	−
Referential Independence	−	−	+

Using these parameters, they suggest that the binding properties of referring devices are a function of predicates, such that the use of a SELF-anaphor (i.e. *zichzelf*) renders the predicate of which it is an argument "reflexive-marked". Simultaneously, they define a "reflexive predicate" as a predicate in which two arguments are co-indexed. For convenience, these definitions are given in (50). These definitions of reflexivity then allow for a re-formulation of binding theory in terms of reflexive predicates and reflexive marking, given in (51). Simply put, condition (a) forces a predicate involving a SELF-anaphor to contain co-indexed arguments. Complementarity, condition (b) requires a predicate with co-indexed arguments to contain a SELF-anaphor (or else be inherently reflexive).

(50) **Reflexivity**

a. *Reflexive predicate:* a predicate P is reflexive iff two of its arguments are coindexed

b. *Reflexive marking:* a predicate P is reflexive marked iff either P is inherently reflexive, or one of P's arguments is a SELF-anaphor

(51) **Binding Theory**

a. A reflexive-marked predicate is reflexive

b. A reflexive predicate is reflexive marked

Together, these conditions neatly capture the the pattern discussed above for Dutch. Consider the fact that only *zichzelf* may be used in simple transitives and thematic prepositional objects (examples (45) and (46) above). In these examples, *hem* and *zich* are ruled out under condition (b): co-indexation with the subject would make the predicate reflexive, but neither *zich* nor *hem* can make it reflexive marked. In contrast, in predicative prepositional phrases, the referring device is an argument of the preposition, not the verb. Consequently, there is no single predicate with

co-indexed arguments, and therefore no reflexive predicate in need of reflexive marking, allowing *zich* and *hem* to be used in sentnces like (47).

However, the (reported) unacceptability of *zichzelf* in (47) proves to be problematic here. As it stands, we might appeal to the principles in (50) and (51) to say that *zichself* reflexive-marks the predicative PP. Since this PP is not, itself, reflexive, condition (a) would then rule the use of *zichzelf* out in these cases. However, as we will see shortly, it is critical for Reinhart and Reuland that reflexive marking does *not* apply to predicative PPs. If so, then we should expect *zichzelf* to be as acceptable as *hem* and *zich* in (47). Notably, this is the only case the authors report in which *zichzelf* is marked ungrammatical in the complement of a predicative PP. All other examples involving predicative PPs report only the judgments for *hem*/*zich*, or else occur in English, where the use of *himself* is perfectly acceptable. Given this, the problematic *zichzelf* judgment may be reported in error, in which case no conflict arises. Independent consultation with two Dutch speakers provides vacillitory evidence on this point: one speaker prefers *zich*, while another prefers *zichzelf*. Interestingly, both disprefer the use of *hem*, but note that all three variants are available, in contrast to the situation in simple transitive clauses (e.g. 45).

Setting aside this case, we can now understand the acceptability of *zich* in inherently reflexive predicates and ditransitives. In the case of inherent reflexivity, co-indexation of *zich* with the subject will result in a reflexive predicate. Since inherently reflexive predicates are automatically reflexive marked, however, condition (b) is satisfied. Likewise, in ditransitives in which both object arguments refer to the subject, reflexive marking is achieved as long as *one* of those object arguments is a SELF-anaphor, leaving the other free for *zich*. Moreover, condition (b) correctly predicts the impossibility of *both* arguments being realized with *zich*, as this would fail to reflexive-mark the predicate.

This leaves two facts unexplained: (1) the variable acceptability of SELF-anaphors in inherently reflexive predicates (or as both object arguments in a ditransitive), and (2) the unacceptability of *hem* in both cases. Reinhart and Reuland explain the first fact with an appeal to an economy principle which disfavors redundantly reflexive-marking an expression. If a predicate is inherently reflexive marked, they suggest, then using a SELF-anaphor would needlessly duplicate the marking[15]. Under the assumption that redundant marking is dispreferred, the relative unacceptability of these sentences is explained.

[15]Likewise, using a SELF anaphor for both object arguments in a ditransitive would "doubly reflexive mark" the ditransitive predicate

The second problem is actually the more difficult[16], and Reinhart and Reuland depart from Binding Theory in explaining the unacceptability of *hem* in (48) and (49). First, they note that their reformulation of binding theory only makes use of one of the parameters they lay out for referring devices: the reflexivizing function associated with SELF-anaphors. Given this, they suggest that *referential independence* is key, and that referentially independent expressions bear the feature +R. Ultimately, then, they propose a revision of chain theory to explain these facts, but before we explore that hypothesis, a bit of exposition is in order.

Reinhart and Reuland, following Chomsky 1973, note that the domain in which an NP can bind its trace is the same in which it can bind a reflexive (−R) but cannot bind an pronoun (+R). They illustrate this with the examples reproduced here in (52) and (53). As can be seen, wherever NP can bind a trace, it can bind a reflexive, but cannot bind a pronoun. In earlier incarnations of the theory (e.g. Chomsky 1973) this intuition was taken as evidence that movement was itself governed by Binding Theory. Reinhart and Reuland would like to suggest that the fundamental intuition was correct, but shift the onus of explanation away from Binding Theory itself, and on to chain theory.

(52) a. Felix$_i$ was fired t_i

b. Felix$_i$ behaved himself$_i$/$*$him$_i$.

c. $*$ Who$_i$ did Felix$_i$ behave t_i?

(53) a. He$_i$ is believed [t_i to be smart].

b. He$_i$ believes [himself$_i$/$*$him$_i$ to be smart].

c. $*$ Who$_i$ does he$_i$ believe [t_i to be smart]?

Building on this parallelism between the binding potential of an NP for traces and anaphora, suggest a revision of Chain Theory (Chomsky, 1986; Rizzi, 1990). Under this theory, an "A-chain[17]" is composed of a sequence of co-indexation headed by an A-position, such that each co-indexed link is c-commanded by another link without an intervening barrier. Given this definition, Reinhart and Reuland suggest the following constraint on A-chains:

(54) **Condition on A-chains:** A maximal A-chain (α_1,...,α_n) contains exactly one link (α_1) that is both +R and case-marked

[16]Inasmuch as one buys their explanation of the first problem. *Caveat emptor.*

[17]Simplifying for the sake of current exposition.

Using this constraint, it is relatively straightforward to explain the divergence between *hem* and *zich*. In inherently reflexive predicates and ditransitives, *hem* would be the tail of a chain headed by the subject. This would violate the chain condition in (54) because now both the head, and the tail of the chain are +R and case-marked. Thus, we have solved problem (2), identified a few paragraphs ago. Handily, this chain condition also explains the unacceptability of anaphors (either SELF or SE) in subject position. If you recall, the definition of Binding Theory given in (51) makes no reference to the fact that the anaphor must be c-commanded by its antecedent. That is, it predicts that sentences like (55) should be acceptable: the predicate is reflexive (two of its arguments are co-indexed), and also reflexive-marked (it contains the SELF-anaphor *himself*). The chain-condition thankfully rules this possibility out by stipulating that a chain must be headed by a +R expression, and SELF-anaphors are crucially −R. Thus, Reinhart and Reuland's chain condition both explains the differing distribution of *hem* and *zich*, and rules out −R expressions (*zich*, *zichzelf*, *himself*, etc.) from c-commanding their antecedents.

(55) ∗ Himself saw John in the mirror.

This brings us to the final (for this discussion) prediction of the Reinhart and Reuland model: when a SELF-anaphor is not an argument of a predicate, it should not enforce local binding. This follows from the fact that a predicate is only "reflexive marked" if one of its *arguments* is a SELF-anaphor. Thus, if a SELF-anaphor is a non-argument, it cannot reflexive-mark a predicate, and thus cannot enforce local binding. Reinhart and Reuland capitalize on this to explain the facts previously observed by Pollard and Sag (1992): when an anaphor is not co-argument with some other antecedent, it is locally free. This explains the fact that reflexives embedded in conjunctions my refer non-locally, as in (56).

(56) Max boasted that the queen invited Lucy and himself for a drink.

However, this raises the specter of an issue raised above: do NPs (or predicative PPs) count as predicates which can be reflexive-marked? If they do, then examples like (57) should be ungrammatical, as the SELF-anaphor would render the local NP/PP "reflexive-marked", but lack a co-indexed antecedent to make the predicate reflexive. Recall that this was the issue identified above in the discussion of predicative PPs in Dutch. A strict reading of Reinhart and Reuland's condition (a) would lead us to expect all such examples to be ungrammatical.

(57) a. The picture of himself that John saw in the post office was ugly.

b. Max saw a ghost next to himself.

Their solution is to suggest a redefinition of predicate, and a refinement of the binding theory, to accomodate these facts. These changes are given in (58) and (59). In brief, these emendations differentiate syntactic predicates, which must include a subject, from semantic predicates, which need not. Concomitantly, condition (a) of their binding theory now only applies to syntactic predicates.

(58) **Predicates**

a. *Syntactic predicate:* the projection containing a head P, all projections assigned a θ-role or case by P, and an external argument (the subject)

b. *Semantic predicate:* the projection containing P, and all its arguments at the relevant semantic level

(59) **Binding Theory**

a. A reflexive-marked syntactic predicate is reflexive

b. A reflexive semantic predicate is reflexive marked

As a result of these changes, condition (a) no longer applies to the NP and predicative PP examples in (57). Crucially NPs (possessorless ones, at any rate) and PPs lack a subject. Therefore, they cannot act as syntactic predicates (as defined by Reinhart and Reuland), and can only be semantic predicates. Consequently, the "predicate" reflexive-marked by a SELF-anaphor embedded in an NP/PP is *not* syntactic, and therefore not subject to the new definition of condition (a)[18]. Thus, Reinhart and Reuland capture much the same facts as Pollard and Sag (1992), albeit with a very different notion of "predicate", and an entirely different formulation of binding theory.

1.2.2.3 A comparison with Pollard and Sag (1992)

Reinhart and Reuland's model, much like Pollard and Sag (1992)'s, is intended to capture two facts: (1) overlap among referring devices, (2) locally unbound uses of reflexive anaphors. Neither of these facts is expected on the SBT model, which predicts strict complementarity among referring devices due to the local-binding requirement of anaphors. Both of the models presented in this section have addressed these challenges to the standard theory by appealing to a predicate-based understanding of binding: local binding is enforced only when two arguments of the same predicate are co-indexed.

[18]Note that the authors further leverage the notion of semantic predicate to explain the distribution of pronouns in sentences like "$*$The queen$_i$ invited both max and her$_i$ to our party". Given that we are less concerned with pronouns in this text, I have omitted this discussion for the sake of concision.

Although fundamentally similar, these two theories go about enforcing local binding in notably different ways. Pollard and Sag suggest that local binding is a function of co-indexed arguments in a subcategorization frame. Reflexive and reciprocal anaphors are required when a more oblique argument is co-indexed with a less oblique argument. In contrast, Reinhart and Reuland present a dual-dissociative system, in which co-argument co-indexation results in a reflexive predicate, which must itself be reflexive marked with a SELF-anapahor (and vice versa). Ultimately, both theories essentially capture cases in which reflexive anaphors are not locally bound. For Pollard and Sag, such cases arise when the anaphor is the highest (or only) referential device on a predicate's subcategorization list. For Reinhart and Reuland, non-local binding occurs when SELF-anaphor is not part of a syntactic predicate.

Furthermore, both theories note that when reflexive anaphors are not locally bound, their use seems to be governed by discourse-oriented, "logophoric" constraints. Pollard and Sag note this similarity, but shy away from labeling "exempt anaphors", as they call them, logophoric. Reinhart and Reuland, on the other hand, note the similarity and, then co-opt the term "logophor" to refer to all situations in which a SELF-anaphor is not part of a syntactic predicate. This mistaken decision has rather confused discussions of exempt anaphors and logophors in the subsequent literation, and will not be recapitulated here. Instead, I will continue to adopt Pollard and Sag's "exempt anaphor" terminology when referring to non-argument reflexives, and reserve "logophor" for the class of West African (and related) pronouns we will be discussing shortly.

Despite the fact that both theories predict the possibility of non-local reference with reflexive anaphors, at least some of their predictions remain incommensurate. To begin with, it isn't clear how Pollard and Sag's theory would handle the case of *zich*. The theory they elaborate is entirely concerned with the distribution of reflexives and reciprocals in English, and the constraints they lay out for these are fairly categorical. Given this, an intermediate pronominal like *zich* poses a problem: in some cases o-commanded *zich* needs to be disallowed (e.g. transitive objects, thematic prepositions), while in others it needs to be preferred (e.g. inherently reflexive predicates, ditransitive objects). On this front, at least, Reinhart and Reuland's theory appears to have the upper hand.

The two theories are on more equal footing with respect to subject anaphors, though they provide decidedly differently explanations. In Pollard and Sag's theory, anaphors may not be the subjects of tensed clauses because they lack the appropriate nominative form, and are thus ruled out on morphological grounds. In contrast, Reinhart and Reuland explain this with their redefinition of the chain constraint, in which $-R$ expressions cannot head A-chains. While different, it

isn't clear that one of these explanations is preferable to the other, and so this does not provide a decisive point in favor of either theory.

However, the tables turn with respect to possessor anaphors. Recall examples like (60), below. Under the standard theory, this example was accounted for by relativizing the size of the binding domain to the referring device under consideration (by incorporating BT-compatibility into the definition of the Minimal Governing Category). In Pollard and Sag's theory, this is a case of exempt anaphor: as the "subject" of the NP, the possessive reciprocal is the highest referring device on its subcategorization list, and therefore free to refer non-locally. However, under Reinhart and Reuland's theory, this example *should* be ungrammatical[19]. Inasmuch as possessed-NPs are "syntactic predicates" on Reinhart and Reulands theory, we should expect the phrase *each other's friends* in (60) to be reflexive marked. Therefore, condition (a) of their theory requires that it also be reflexive. However, this phrase lacks a co-argument with which *each other* could be co-indexed, leading us to expect that, under the Reinhart and Reuland model, this sentence should be unacceptable.

(60) The boys like each other's friends.

Finally, the two theories handle ECM constructions entirely differently, and with variable success. That is, they afford different explanations for the reflexives in sentences like (61). On their face, these examples *should* be counter evidence to both theories: the critical reflexive is the subject of an embedded infinitaval/small clause, and therefore not co-argument with the matrix subject. However, in Pollard and Sag's account, this is not the case. Critically, in HPSG (and related theories), the anaphor in these cases is taken as the primary object of the matrix verb, rather than an embedded clause subject. As a result, it sits as the second argument of the matrix subcategorization frame, and thus may be o-commanded by the matrix subject.

(61) a. Mary believes herself to be superior.

b. They consider themselves superior.

c. We regard each other as imposters.

In contrast, ECM structures prove slightly more difficult for Reinhart and Reuland. The SELF-anaphor in these examples is assigned case by the matrix verb, and so is (appropriately) part of the matrix predicate. This results in the correct prediction that the embedded subject is obligato-

[19]Assuming a parity between reflexives and reciprocals. Notably, Reinhart and Reuald assiduously avoid their reciprocals in their discussion, except to speculate that they should behave analogously with reflexives

rily bound by the matrix subject. However, the embedded subject is *also* in the thematic domain of the embedded predicate, which is consequently also reflexive marked. Given that the embedded predicate is not reflexive (under Reinhart and Reuland's definition) this would predict that sentences like (61) should be ungrammatical. Reinhart and Reuland patch this problem by invoking a verb raising operation which adjoins the embedded verb to the matrix verb. They argue that this operation bleeds the thematic assignment associated with the embedded predicate, and so prevents the condition (a) violation that would otherwise results. Put differently, binding into ECM clauses does not fall out naturally from Reinhart and Reuland's theory, but must be explained via appeal to other syntactic operations[20].

In sum, both theories attempt to address the empirical inadequacies of the SBT model by suggesting that non-argument (henceforth, "exempt") reflexives may find their antecedent non-locally. The theory advanced in (Pollard & Sag, 1992) is better equipped to deal with possessor anaphors and ECM constructions, but does not (in its current form) capture the facts for *zich*, which is neatly explained in Reinhart and Reuland's theory. However, neither of these theories predicts the contrast in (inanimate-anaphors), repeated here in (inanimate-anaphors2), a theme to which we shall return in Section 1.4.

(62) a. John's email suggested that opinions about himself would be divided.

b. $*$ The email's subject line suggested that opinions abut itself would be divided.

1.2.3 Binding Theory: a summary

So far in this chapter we have seen three different interpretations of the constraints which hold of binding for reflexives and pronominals. It is, I think, at this point useful to evaluate the dimensions on which these theories present similar predictions, as well as those on which they differ, as well as briefly reminding ourselves of their central components.

As we saw in Section 1.1, the Standard Binding Theory (SBT) is built around the central intuition that the referential capabilities of anaphors (reflexives and reciprocals) and pronominals are in complementary distribution. In accounting for this distribution, the SBT model defined locality in such a way as to ensure that anaphors referred locally, while pronominals never did (Chomsky, 1986). In contrast, Section 1.2 has presented an alternative approach in which complementary distribution was *not* the starting assumption. Instead, these theories sought to explain binding in

[20]Notably, these examples are straightforwardly captured in the SBT model, where the embedded clause in (61) cannot be the MGC for the anaphor.

terms of co-arguments and reflexive predicates. Reflexivity is enforced when multiple arguments of a predicate are co-indexed, and not otherwise. SBT models struggle with non-local interpretations of reflexives, while predicate-based models accommodate ECM (and similar) constructions with variable success. Only Reinhart and Reuland's model directly accounts for the distribution of *zich*.

In spite of these different implementations, all three theories predict that direct-object anaphors should necessarily be locally bound. In other words, regardless of your formulation of Principle A, a direct object reflexive must be bound by its local subject. This will provide the basis for the sentence processing investigations presented in subsequent sections.

1.3 Reflexive pronouns in sentence comprehension

Stepping back from theory, for a moment, we turn to an overview of reflexive pronouns in the sentence comprehension. The focus of this section will, consequently, be fairly different. Rather than considering various approaches to explaining the distribution of referring devices, we will be more concerned with understanding how binding theory constraints are applied in real time. As a starting premise, lets assume that at the point a referring device is encountered, the intended referent is not immediately, and automatically available, and some mechanism must be employed to search memory and identify an appropriate antecedent. How, then, does binding theory impact this search process? Do the various theoretical models discussed above suggest different predictions for how this search process should unfold?

With respect to this second question, at least, the current section will be unable to provide a meaningful answer. Perhaps unsurprisingly, investigations of reflexive pronouns in the sentence processing literature have largely avoided the varied, and nuanced structures discussed in previous sections. Instead, this literature has largely focused on reflexives in direct object positions, with occasional forays into the exotic territory of non-argument reflexives[21]. In so doing, such investigations typically assume a much simplified version of binding theory, closest in spirit, perhaps, to the definition in (3) of the Standard Binding Theory. For expository purposes, I give a general formulation of binding theory as it is typically discussed in the sentence processing literature in (63), below. Reference to "Binding Theory" in this section, then, intends this interpretation.

(63) **Binding Theory:**

[21] In other words, the cases in which all three theories make the same predictions.

a. A reflexive must be bound in the local clause

b. A pronoun must not be bound in the local clause

Slightly more headway can be made in answering the question of how Binding Theory impacts antecedent identification, however. To get ourselves started, lets consider two dimensions on which binding theory might impact this process: time, and strength. With respect time, Binding Theory could act as a direct, and immediate filter on the search process, constraining the search space to include only those candidates which would be compatible with the grammar. Conversely, Binding Theory could act as a late filter, allowing a very broad search space to eventually be winnowed down to those few candidates which satisfy its constraints. Thus, we might have a very early, or very late, application of Binding Theory in comprehension. In terms of strength, Binding Theory could act as a rigid constraint on antecedent identification, in which case infelicitous antecedents are never considered (or always rejected, depending on the timing). Alternatively, it might only provide weak guidelines for identifying antecedents, allowing sub-par referents to be selected at the comprehenders' whim. Logically, this produces four possible pictures of binding theory in sentence comprehension[22], represented here in Table 1.2.

Table 1.2. The logical space of models of Binding Theory in sentence comprehension given the dimensions of time, and strength of application

	Timing	
Strength	*early, strong*	*late, strong*
	early, weak	*late, weak*

In the case of an early, strong application of Binding Theory, we expect to find evidence that there is no point at which comprehenders consider referents inconsistent with binding principles. In contrast, a late, strong model would predict that at early stages of comprehension, comprehenders may attend to illicit referents, but later strongly reject these interpretations. An early, weak model might be represent Binding Theory as a "defeasible filter" on interpretations—comprehenders initially only consider Binding Theory compatible antecedents, but may adopt other interpretations at a later stage. Finally, a late, weak model represents the least influence of binding constraints on antecedent identification: all referents are under consideration, and Binding Theory only weakly applies at a late stage to privilege some antecedents over others. In what follows, I present evidence from the sentence processing literature that Binding Theory is applied

[22] I suppose an infinite number of alternative models, in which both time, and strength are more gradiently defined, are hypothetically possible. However, models which apply Binding theory at 30ms after encountering a referring device with 63% strength do not exactly lend themselves as natural starting points for our investigation. For the present, then, I will remain simple, and binarily minded.

early in the course of antecedent identification, but with somewhat variable strength. While this will, occasionally, involve discussion of pronominals, our primary focus will remain with reflexive pronouns and Principle A.

1.3.1 Evidence of an early application of Binding Theory

One early attempt at addressing the real-time application of Binding Theory comes from Nicol and Swinney (1989), who investigated the question in a series of cross-modal semantic priming experiments. Sentences like (64) were presented to participants auditorily. At the offset of the anaphor/pronoun, the recording was interrupted and participants were asked to perform a lexical decision task for a word presented on a computer screen (i.e. "is this word of English?"). Critically, this word was manipulated so that it was either a semantic associate of one of the three preceding referents (e.g. *gloves* for *boxer*, *lift* for *skier*, or *nurse* for *doctor*), or a completely unrelated word (e.g. "food"). Of interest, then, is whether participants are *faster* to respond to words related to the preceding referents, as this would indicate that these referents have recently been reaccessed as a result of processing the referring device.

(64) The boxer told the skier that the doctor for the team would blame $\left\{\begin{matrix}\text{himself}\\ \text{him}\end{matrix}\right\}$ for the injury.

If the (simplified) version of binding theory sketched above is being applied early to constrain the search space for antecedents, we expect to find semantic priming for associates of *doctor* when the referring device was a reflexive, but not for either *boxer* or *skier*, as these referents are outside the local clause. In contrast, if the referring device was a pronoun, we expect semantic priming for associates of *boxer* and *skier*, but not for associates of *doctor* (since pronouns need to be locally free). This is precisely what Nicol and Swinney (1989) report. When the referring device was a reflexive, they observed substantial semantic priming for local antecedents, but not for distal ones; when it was a pronoun, associates of distal referents were primed, but not local ones[23]. For reference, these data are reproduced in Table 1.3.

Studies like this provide suggestive evidence that referents incompatible with Binding Theory are not reactivated in the course of antecedent resolution. However, these findings do not rule out a (relatively) late application of Binding Theory. That is, comprehenders might access all potential antecedents, and then apply Binding Theory to inhibit those which are infelicitous. This would

[23]In a case of spooky symmetry, priming for local associates after anaphors was essentially double that observed for distal associates after pronouns. Moreover, priming for distal associates was nearly evenly split among the two referents. Really, it's almost enough to make one believe in spreading activation.

Table 1.3. Semantic priming data from Nicol and Swinney (1989) (unrelated−related). Positive values represent facilitation due to semantic priming (∗ indicates significance)

	Referring Device	
Referent	*himself*	*him*
Boxer	-1	43∗
Skier	11	58∗
Doctor	104∗	-21

produce the same pattern of behavior, but result from a model in which Binding Theory did not *initially* constrain the search process.

Given this ambiguity, more recent approaches (starting with Badecker & Straub, 2002) have employed a slightly different method of assessing the impact of Binding Theory on sentence comprehension. In these studies, researchers measure the impact of Binding Theory (in)compatible referents by manipulating the feature-match of referring devices and their putative referents. This "mismatch" paradigm thus allows for an independent assessment of the impact of antecedents ruled-in by Binding Theory (TARGETS), and those incompatible with binding constraints (LURES). A sample stimulus from this paradigm is given in (65). Given a reflexive anaphor, a local, c-commanding referent constitutes a TARGET, while a non-local (or non-c-commanding) referent is a LURE. In (65), the gender of each of these antecedents is manipulated such that they either match, or mismatch the embedded reflexive. To the extent that we observe a difference in behavior between target match, and mismatch conditions (a "target match" effect), we have evidence that comprehenders are attending to the features of antecedents compatible with Binding Theory when processing the reflexive. Likewise, finding differences between lure match and mismatch conditions is an indication that comprehenders are attending to the features of lures, even though Binding Theory rules these referents out as potential antecedents.

(65) $\left\{\begin{matrix}\underline{\text{LURE}}\\ \text{Jonathan}\\ \text{Jennifer}\end{matrix}\right\}$ saw that the $\left\{\begin{matrix}\underline{\text{TARGET}}\\ \text{schoolboy}\\ \text{schoolgirl}\end{matrix}\right\}$ had hurt himself...

Perhaps the first study to employ this paradigm was Badecker and Straub (2002), who investigated sentences like (66) in a word-by-word self-paced reading study. This study did not manipulate the target, but did manipulate the lure, providing an index of attention to lures when reading reflexive pronouns. Somewhat surprisingly, Badecker and Straub report a "multiple match" effect, such that reading times were slowed when both lure, and target matched the reflexive (relative to when only the target matched). On its face, this would seem to be evidence that comprehenders do, at some stage, consider antecedents incompatible with Binding Theory, and therefore support either a weak, or late application of binding principles. However, there are two points which com-

plicate this interpretation. First, the "multiple match" effect Badecker and Straub report emerged only at a delay from the critical reflexive, a full two-words later than the reflexive itself. This suggests that there may be an early application of Binding Theory which is only defeasible at a later stage (i.e. a "early weak" model, rather than a "late" one). Second, the authors failed to replicate this effect in a subsequent study in which the lure did not c-command the reflexive, suggesting that the effect may not be stable. This concern is compounded by the fact that replication of the "multiple match" effect has proven elusive in subsequent work, with only one study managing to replicate it (Patil et al., 2016) to date[24].

(66) $\left\{\begin{matrix}\text{Jane}\\\text{John}\end{matrix}\right\}$ thought that Bill owed himself another opportunity to solve.

Directly relevant to both of these concerns, Sturt (2003b) investigated the time course and replicability of Badecker and Straub (2002)'s multiple match effect a series of influential eye-tracking while reading studies. In separate experiments, he investigated sentences like (67)[25]. Notably, Sturt found strong effects of target match: reflexives which mismatched the target antecedent were read more slowly than those which matched the target antecedent. However, in early measures of reading difficulty he found no evidence of a lure match effect, indicating that participants attend to the features of target antecedents quite early, but ignore those of lures. It's worth noting that Sturt did observe lure-match effects in later measures of reading difficulty (second-pass reading time), but that these effects did not replicate Badecker and Straub's multiple match effect, and have themselves failed to replicate in later studies.

(67) Jonathan/Jennifer was pretty worried at the city hospital.
$\left\{\begin{matrix}\text{He}\\\text{She}\end{matrix}\right\}$ remembered that the surgeon had pricked $\left\{\begin{matrix}\text{himself}\\\text{herself}\end{matrix}\right\}$ with a used syringe needle.
There should be an investigation soon.

Subsequent studies using variations on Sturt's mismatch paradigm have produced similar findings: in an EEG experiment, Xiang et al. (2009) found no effects of lure match on early ERP components; and in another eye-tracking while reading study which manipulated number, rather than gender, Dillon et al. (2013) likewise found no effects of lure match on early measures of reading difficulty at the reflexive. Perhaps most convincingly, a recent meta-analysis of sixteen

[24]In English. There is suggestive evidence that the multiple match effect may exist in Mandarin (Chen et al., 2012), but this data suffers from many of the same concerns as apply to Badecker and Straub. In particular, both studies used self-paced reading, and the critical effects were observed on post-reflexive material, rather than on the reflexive itself. As Sturt (2003b) notes, this pattern of results is consistent with an early, weak model in which Binding Theory is applied as an initial, defeasible filter on antecedent identification.

[25]While superficially different from the example in (65), Sturt (2003b)'s paradigm is essentially the same. The difference is that Sturt manipulated target match by varying the critical reflexive, rather than the target antecedent itself.

separate experiments on reflexive comprehension in English[26] found only sparse evidence of lure-match effects on reflexive comprehension (Jäger et al., 2017). In target-mismatch conditions, this analysis reports a significant lure-match effect for only one study (out of a possible eleven). In this case (Cunnings & Sturt, 2014), feature-matched lures produced a slow-down in reading time, albeit at a delay (the effect only emerged at the post-reflexive region). Thus, this finding is consistent with the view that Principle A is applied early, but defeasibly.

In target match conditions, Jäger et al. (2017)'s meta analysis finds two significant lure-match effects (of a possible sixteen). One of these is Badecker and Straub (2002)'s multiple match penalty effect, discussed above. The other was an effect in the *opposite direction*, in which feature-matched lures actually facilitated reading times of the reflexive ("lure-match facilitation" Cunnings & Felser, 2013). Interpretation of this finding is further complicated by the fact that it was only observed for "low span" readers, suggesting that the behavior may be tied to reading behaviors adopted by non-expert readers.

In sum, there has been relatively scant evidence that lure referents impact the early stages of antecedent identification[27]. To the extent that lures *do* have an impact, their effect seems to be somewhat variable (helping, and impeding in equal measure), and at a delay from the initial stages of antecedent identification[28]. These findings have led some researchers to conclude that the early stages of reflexive processing are tightly constrained by Principle A, leading the parser to categorically ignore lure referents (Sturt, 2003b; Xiang et al., 2009; Dillon et al., 2013, i.a.) .

1.3.2 Evidence of a weak application of Binding Theory

Thus far, this section has presented a fairly straightforward picture: reflexive binding is a function of locality and c-command, and the parser seems to directly apply these constraints in resolving anaphoric reference. Unfortunately, this simple picture doesn't quite hold. For one thing, much of the previous section was devoted to a discussion of why locality and c-command are insufficient (on their own) as an explanation of binding behavior (Pollard & Sag, 1992; Reinhart & Reuland, 1993; Charnavel & Sportiche, 2016). However, even setting this concern aside, recent evidence from sentence processing suggests that comprehenders do, occasionally, access lure referents in the course of antecedent resolution. In two eye-tracking while reading studies, Parker

[26]Jäger et al. (2017)'s meta analysis also included three experiments on *ziji* in Mandarin, and two experiments on reciprocals. Given our focus on English reflexives in this section, I do not discuss those results here.

[27]For reflexives. We have not seen much data in this section relevant to the same question for pronouns.

[28]For a comprehensive review, see (Dillon, 2014; Sturt, 2013; Jäger et al., 2017)

and Phillips (2017) investigated sentences like (68), manipulating the degree of feature-mismatch between an embedded reflexive and its target antecedent. Congruent with previous findings, he found no effects of lure match when the reflexive matched the target antecedent (*boy*), or when it only mismatched in a single feature (*girl*). However, when the reflexive and target mismatched in two features (*girls*), the target-match penalty was ameliorated by the presence of a feature-matched lure. Based on this finding, Parker and Phillips concluded that reflexive processing is sensitive to lure referents only when the target antecedent is an exceptionally poor match for the reflexive's phi-features.

(68) $\overset{\text{LURE}}{\left\{\begin{matrix}\text{Steven}\\ \text{Susan}\end{matrix}\right\}}$ said that the $\overset{\text{TARGET}}{\left\{\begin{matrix}\text{boy}\\ \text{girl}\\ \text{girls}\end{matrix}\right\}}$ embarrassed himself...

Findings like Parker's present a challenge for the simple view sketched above. Principle A (at least as formulated in 63) isn't an absolute check on reflexive reference in online comprehension, nor could these effects be explained by appealing to alternative versions of the Binding Theory. Recall that all three variants we considered above made the same predictions for argument reflexives—they should be obligatorily bound. Thus, under any of the variations considered so far, an early, strong application of Principle A would predict no sensitivity to lure referents in Parker and Phillips studies. Notably, these finding still don't indicate a late application of Principle A: when target antecedents are only a relatively poor morphsyntactic match, Binding Theory still seems to tightly constrain the antecedents under consideration. Thus, these findings locate the impact of Binding Theory on antecedent identification somewhere in the "early, weak" category: Principle A is immediately deployed to constrain the search process, but in an imperfect, and decidedly fallible manner. We turn now to a consideration of the mechanisms which might instantiate such a system.

1.3.3 Cue-based parsing models

Findings which indicate that comprehenders consider Principle A incompatible referents in on-line processing have been used to argue for cue-based parsing models (Chen et al., 2012; Parker, 2014; King, Andrews, & Wagers, 2012; Cunnings & Felser, 2013; Patil et al., 2016, i.a.). Specific proposals of such models include the ACT-R memory architecture of Lewis and Vasishth (2005), the augmented ACT-R system proposed by Engelmann, Jäger, and Vasishth (Submitted), and the content-addressable, cue-based retrieval mechanisms proposed by Lewis et al. (2006); McElree (2000, 2006); McElree, Foraker, and Dyer (2003); Van Dyke (2007); Van Dyke and Lewis (2003), and Van Dyke and McElree (2006). All of these models share in common the proposition

that parsing operations (e.g. structure building, attachment decisions, referential access, etc.) are subserved by a single, content-addressable retrieval mechanism. This mechanism retrieves objects from memory to resolve open linguistic dependencies using a set of features (retrieval cues) to activate stored representations. To accommodate a wide variety of dependencies, these cues vary as a function of the particular dependency represented in the currently attended input. Importantly, this mechanism is sensitive to, but may not uniquely privilege, syntactic constraints on dependency formation, placing such information on par with surface cues to dependency resolution (but see (Van Dyke & McElree, 2011) for discussion of the role of syntactic constraints on retrieval processes).

In the case of reflexives, these models hypothesize that comprehenders attempt to retrieve an antecedent by querying memory for referents in an appropriate structural position, with appropriate morphosyntactic features. Since this retrieval process is stochastic and error prone, it may accidentally select an unintended (i.e. lure) referent if its morphosyntactic features match the retrieval cues, thus allowing these models to account for both multiple match (Badecker & Straub, 2002; Chen et al., 2012; Patil et al., 2016) and facilitatory interference effects (King et al., 2012; Parker, 2014). In the first case, multiple match effects arise when multiple referents match the morphosyntactic cues engaged at a reflexive to search for an antecedent. Because the target antecedent is no longer uniquely singled out by the search cues, it is more difficult to retrieve, a phenomenon known as similarity-based interference. Similarly, facilitatory interference arises when the target antecedent does not match the morphosyntactic features of the reflexive, but the lure referent does. In this case, the retrieval process is liable to retrieve the feature-matched lure, giving rise to the (at least temporary) percept of well-formedness (the illusion of grammaticality).

Thus, cue-based parsing models provide a reasonably straightforward way of modeling an early, weak implementation of Binding Theory in antecedent retrieval. The early impact of Binding Theory derives from the fact that c-command and locality are built in as search cues for retrieval operations. Therefore, the model will be strongly biased to retrieve antecedents which satisfy binding constraints. Binding Theory's inability to *completely* constrain behavior, then, arises from the necessarily stochastic nature of the retrieval mechanism. Since this retrieval is also cued with morphosyntactic features, an unintended lure referent will, on occasion, be reaccessed in contravention of Principle A.

That said, cue-based parsing accounts of Principle A fallibility face in turn a number of thorny implementational questions. For example, how are c-command and locality encoded in these kinds of models? Both notions are relational, and require a fully specified parse to accurately artic-

ulate, raising the question of how best to dynamically mark constituents with locality/command features in an incremental parser[29]. Questions of implementation aside, such models leave as a puzzle the striking divergence between reflexive comprehension and other, surface-similar dependencies. Error-prone retrieval has been widely adopted as an explanation of agreement attraction, a phenomenon which is readily observed and has been repeatedly documented (Wagers, Lau, & Phillips, 2009; Eberhard, Cutting, & Bock, 2005, i.m.a.). In contrast, interference in reflexive dependencies has proved relatively elusive, despite the fact that these dependencies engage surface-similar phi-features. While this contrast could be accommodated in a cue-based framework, the source of the divergence remains something of a mystery. Put differently, there's an independent question of whether cue-based models provide a satisfying explanation of variation in Principle A fallible behavior. Is sensitivity to lure referents entirely a function of feature match and activation decay (i.e. is grammatical fallibility the product of probabilistic retrieval error?), or are there other factors at play, possibly overlooked in prior manipulations?

Put slightly differently, the finding that lure referents are available in online comprehension is compatible with the view that the parser entertains ungrammatical interpretations during comprehension as a consequence of the memory access routines engaged in identifying an antecedent (Van Dyke, 2007). However, this conclusion necessarily assumes that taking lure referents as antecedents in these studies is, in fact, *ungrammatical*. That is, that no licensed or conventional interpretation of the reflexive form permits co-reference with Principle A incompatible referents (at least for speakers of standard American or British English). In the present work, I would like question this assumption. While these interpretations are ill-formed with respect to conventional versions of the Binding Theory (Chomsky, 1986; Pollard & Sag, 1992; Reinhart & Reuland, 1993), it remains possible that other grammatical principles (albeit non-Binding-Theoretic ones) allow reflexive forms to co-refer with 'lure' referents. Put differently, findings like Parker and Phillips (2017) may not represent situations in which reflexive comprehension must arbitrate between grammatically permissible targets, and unacceptable lures, but rather situations in which multiple possible *targets* compete for reflexive reference. Taking this perspective, reflexive "illusions of grammaticality" are not *illusions* at all, but rather reflect an alternative, grammatical interpretation of the reflexive form capable of referring outside the scope of Principle A. This interpretation

[29]However, see recent work (e.g. Kush, 2013) which has gone some way to addressing this issue. Furthermore, this may be a point on which predicate-based theories have an advantage: marking arguments as belonging to a local predicate (and then instructing anaphors to search for other members of that predicate) seems, *a priori*, simpler. I set these considerations aside, for now, and revisit them in greater detail in Chapter 5.

would necessarily lie outside the domain of any of the theoretical binding models we have considered thus far, as all three predict that direct object reflexives should be obligatorily locally bound. Given this, we turn next to a re-evaluation of the versions of Binding Theory presented earlier to see if more modern approaches might provide insight into which other grammaticaly factors (if any) might be in play.

1.4 Principle A fallibility: a role for logophoricity

As noted above, Parker and Phillips (2017) present findings that comprehenders entertain antecedents incompatible with any of the three versions of Binding Theory considered so far in this chapter. However, at least two of these theories (Pollard & Sag, 1992; Reinhart & Reuland, 1993) admit that reflexives refer non-locally in select cases, and that when they do their interpretation is constrained by alternative, discourse constraints (e.g. perspective and focus). Perhaps, then, it is this alternative use of reflexive forms which was being accessed by comprehenders in Parker and Phillips' study. This view would gain further support if we had evidence that such uses were not *grammatically* constrained to non-argument positions, but rather preferentially associated with those positions. Descriptively, this would constitute a model in which SBT was fundamentally correct in its characterization of reflexive anaphors, but failed to account for an alternative, *non*-anaphoric[30] use of the reflexive form. One such theory may be found in Charnavel and Sportiche (2016), whose primary arguments and model I present in brief here.

Following the observations of Pollard and Sag (1992) and Reinhart and Reuland (1993), Charnavel and Sportiche (2016) take seriously the challenges posed by non-local uses of reflexive anaphors. However, they point out that these previous theories failed to take into account differences in the distribution of animate, and inanimate anaphors. That is, under either predicate-based theory, we should expect non-argument, inanimate anaphors to have the same distribution as their animate counterparts—the relevant dimension is syntactic position (argument vs. non-argument), and not animacy. However, at least in the case of French anaphors, Charnavel and Sportice claim that animates, but not inanimates, may refer non-locally. This is shown in the contrast between (69a) and (69b) for the possessive anaphors *son propre* and *sa propre* (*it's*/*his* own).

(69) a. Ce pont$_i$ a l'aire très fragile. Son$_i$ (∗propre) architecte a reçu moins de moyens que les autres architectes de la région.

[30]Interpreting "anaphor" here in the sense of Chomsky (1986) as meaning those referring devices which must be locally bound.

This bridge$_i$ looks very fragile. Its$_i$ (∗own) architect got less means than the other architects of the area.

b. Cet enfant$_i$ a l'aire très perturbé. Sa$_i$ (propre) mère passe moins de temps à la maison que les autres mères de la classe.
This child$_i$ looks very disturbed. His$_i$ (own) mother spends less time at home than the other mothers of the children in the class.

As seen in (69), when the referring device is animate, either the possessive pronominal (*sa*), or the possessive anaphor (*sa propre*) may be used to refer cross-sententially. However, if the referring device is inanimate, only the pronominal (*son*), not the anaphor (*son propre*) may be used. They make a similar point for *elle-même*, a form which does not, itself encode animacy, but the referential capabilities of which vary as a function of the animacy of the intended antecedent, as seen in (70)-(72). In (70), we see that when *elle-même* is locally bound, it is obligatory (even though its antecedent is inanimate). However, when referring non-locally, either the pronominal *elle*, or the reflexive *elle-même* may be used if the antecedent is animate (72), but *elle-même* may *not* be used if it is inanimate (71).

(70) a. ∗ La Terre$_i$ tourne autour d'elle$_i$
The earth revolves around it

b. La Terre$_i$ tourne autour d'elle-même$_i$
The earth revolves around itself

(71) a. La Terre$_i$ subit l'effet gravitationnel des nombreaux satellites qui tournent autour d'elle$_i$
The earth$_i$ is subject to the gravitational effect of the numerous satellites that revolve around it$_i$.

b. ∗ La Terre$_i$ subit l'effet gravitationnel des nombreaux satellites qui tournent autour d'elle-même$_i$
The earth is subject to the gravitational effect of the numerous satellites that revolve around itself.

(72) a. De son point de vue, Marie$_i$ souffre de la présence des nombreuses personnes qui tournent autour d'elle$_i$.
From her point of view, Marie$_i$ suffers from the presence of many people who move around her$_i$.

b. De son point de vue, Marie$_i$ souffre de la présence des nombreuses personnes qui tournent autour d'elle-même$_i$.

From her point of view, Marie$_i$ suffers from the presence of many people who move around herself$_i$.

In light of this, it seems that the inanimate reflexive forms *son propre* and *elle-même* behave unexceptionally like locally-bound anaphors in accord with the predictions of the SBT model. in contrast, only animate anaphors seem to allow non-local reference. This distribution does not fall out naturally from predicate-based theories of binding (Pollard & Sag, 1992; Reinhart & Reuland, 1993), according to which syntactic position is the (primary) relevant factor for determining an anaphor's ability to refer non-locally. At the very least, both of the theories discussed previously would need to incorporate new constraints to explain differences between animate and inanimate anaphors in these environments.

Charnavel and Sportiche suggest a different approach: inanimate reflexives show the true distribution of anaphors, and non-local interpretations of animate reflexives demonstrate an alternative, non-Binding Theoretic use of reflexive forms. Specifically, following Charnavel and Zlogar (2015), they suggest that non-local interpretations are the product of a *logophoric* interpretation of reflexive pronouns. This interpretation is mediated by a local, hidden operator OP$_{\text{LOG}}$ which locally-binds a reflexive form, and itself refers to the "perspective holder" of the utterance (more on that later). This system neatly captures the facts for French, and can be readily extended to English as we shall see shortly. First, it maintains the fundamental generalization made by the SBT model: anaphors are always locally bound. In the case of inanimate anaphors, this is always by a c-commanding referent within the local domain. Animate anaphors, however, may optionally be bound by the local OP$_{\text{LOG}}$, and thereby "refer non-locally" to the utterance's perspective holder. The difference between animate and inanimate anaphors arises from the nature of OP$_{\text{LOG}}$: since this operator's antecedent is the *perspective* holder, it is incompatible with antecedents/anaphors which are inanimate, and therefore incapable of taking a perspective.

Unfortunately, this does not explain why animate anaphors only achieve non-local reference when in non-argument positions. That is, if non-local reference is always, in fact, locally mediated, why do argument reflexives seem to be incapable of finding a binder in OP$_{\text{LOG}}$? Charnavel and Sportiche appeal to an independent constraint on the choice of referring device proposed by Cardinaletti and Starke (1999): all else equal, weaker forms (e.g. clitics) block the use of stronger forms (e.g. reflexive anaphors), if available. In this case, we see a dispereference for using argument reflexives to refer non-locally because a clitic of the appropriate form could be used instead.

This is shown in (73). In contrast, where a clitic may not be used, non-local reference with the reflexive form is acceptable, as in (74).

(73) a. $*$ Jean$_i$ pense que Marie examinera lui-même$_i$.
John$_i$ thinks that Marie will examine himself$_i$.

b. Jean$_i$ pense que Marie l$_i$'examinera.
John$_i$ thinks that Marie will examine him$_i$.

(74) a. Marie$_i$ s'inquiète du fait que ses enfants dépendent d'elle-même$_i$.
Marie$_i$ is worried that her children depend on herself$_i$.

b. $*$ Marie$_i$ s'inquiète du fait que ses enfants la$_i$ dépendent.
Marie$_i$ is worried that her children depend on her$_i$.

However, it isn't clear that this explanation would extend to English, in which no weaker form exists to block the use of long-distance reflexives in argument positions. Here, Charnavel and Sportice follow Ahn (2015) in noting that argument reflexives in English typically do not bear phrasal stress, and so are deaccented. However, reflexives in non-argument positions (e.g. in prepositional complements) *may* bear phrasal stress. Building on this they suggest that long-distance interpretations may be preferentially associated with the phrasal-accented variant of the reflexive, in a manner similar to the clitic/anaphor distinction observed in French. For the present, I take this as the appropriate description for English, though we will return to a discussion of the role played by co-argumenthood in chapter 5.

Setting the issue of how best to capture the intuition that argument/non-argument reflexives behave differently, Charnavel and Sportiche (2016) provide us with a reasonable grammatically-based alternative to Parker and Phillips (2017)'s account: findings of lure sensitivity do not represent *errors* of an underlying memory retrieval mechanism, but rather reflect an alternative, logophoric interpretation associated with reflexive anaphors. To explore this further, then, we need a clearer notion of what "logophoricity" entails.

1.4.1 Logophoric Pronouns

Logophors are pronouns which necessarily refer to the person whose speech, thoughts, or feelings are reported in an utterance (Clements, 1975). Quite generally, they refer to the "perspective holder" (or "perspective center") of an utterance, though languages vary with respect to which aspects of perspective are relevant (Sells, 1987). Most commonly, logophors refer to the "speaker" of an utterance, where the speaker may be identified either with the actual, utterance speaker (i.e.

"first person"), or with some third-person referent whose speech is being reported/represented in a clause(Sells, 1987; Culy, 1994; Speas & Tenny, 2003; Koopman & Sportiche, 1989). Logophoric pronouns (and their agreement variants) are most commonly found in West African languages like Ewe, Abe, and Tupuri (Culy, 1994, 1997), where they are realized with forms which are morphologically distinct from other pronouns and referring devices. However, several authors have argued that long-distance reflexives are, themselves, expressions of a logophoric meaning, suggesting that reflexivity and logophoricity may occasionally share the same morphological form. This claim has been made most strongly for the Mandarin reflexive *ziji* (Huang & Liu, 2001; Anand, 2006), the Japanese reflexive *zibun* (Kuno, 1986; Sells, 1987), and Icelandic (Maling, 1984; Sells, 1987). In this sense, the claims advanced for French by Charnavel and Sportiche (2016), and for English by Charnavel and Zlogar (2015); Loss (2014), and myself, join a long tradition of analyzing non-local anaphora in terms of logophoricity.

With respect to the mechanics of logophoricity, several technical accounts have been advanced. Sells (1987) advances a model couched in discourse-representation theory, suggesting that logophoric pronouns refer to one of a variety of roles represented in the discourse structure and assigned by particular verbs. Similarly, (Culy, 1994) suggests that particular verb classes are responsible for embedding logophoric propositions. Both of these accounts are given more attention in Chapter 2, where the impact of embedding verbs wil be of particular importance. More modern approaches have favored binding logophoric pronouns with operators located in the left-periphery of clauses (Speas & Tenny, 2003; Anand, 2006; Charnavel & Zlogar, 2015). Speas and Tenny (2003), proposed that the left periphery contained a "speaker" operator, which served as the antecedent for logophoric pronouns. Following Koopman and Sportiche (1989), Anand (2006) suggests that attitude verbs may optionally embed logophoric operators. These operators bind embedded logophoric pronouns, and themselves refer to the *de se* center of an alternative world (i.e. the speaker or addressee of that alternative world). In the spirit of Sells (1987), Charnavel and Zlogar (2015) posit a series of logophoric operators corresponding to different levels of discourse role. These operators are ordered with respect to each other in the left-periphery, making binding from some operators more economical (i.e. more local) than others. In all three of these latter accounts, the mechanism by which the operator finds its referent is left underspecified, with the assumption that *either* the most local possible referent binds the operator (Anand, 2006), or else that the operator finds its referent from a discourse model of some kind (Charnavel & Zlogar, 2015).

It is important to note that this notion of "logophoricity" is markedly different from how it is sometimes referred to in both the syntax, and sentence processing literature. For example, Reinhart and Reuland (1993) label all instances of reflexives in syntactically exempt positions "logophoric". This may be the case, but it has served to confuse the issue in subsequent discussions, such that the term "logophoric pronoun" is sometimes used to mean a reflexive pronoun in a non-argument position, and sometimes used to mean the class of pronouns which refer to perspective holders. Usually, the former intends something like the latter, but without making explicit reference to constraints which are known to hold of true, logophoric pronouns. This confusion is especially pernicious in discussions of long-distance anaphora, where it tempting to conflate the "logophoricity" of West African languages with the "logophoricity" of exempt anaphors. The two may yet be related (and indeed, I present evidence that they are), but we should be careful to distinguish non-local reference dependent on syntactic position from true logophoricity. In the event that syntactic dependence is *not* a relevant organizing dimension, then, this distinction may yet be collapsed.

One area in which this distinction has been repeatedly obfuscated is in sentence processing approaches to "logophoricity" (Cunnings & Sturt, 2014; Burkhardt, 2005), where no research on true (i.e. morphologically distinct) logophors has been conducted[31]. For example, (Burkhardt, 2005) reports that processing logophors may be more costly. However, the definition of "logophor" assumed is that of "exempt anaphor", and the sentences investigated include structures like (75), where binding is still plausibly local, albeit into a non-argument position. Thus, it isn't clear that such examples represent investigations of true logophoricity, rather than "exempt" syntactic positions. At present, I set aside further consideration of the processing of "exempt" structures, though they will be discussed in more detail in Chapter 5, when the issue of understanding exemption is re-opened.

(75) Several coworkers heard that Jenny had criticized both herself and Nathan.

1.4.2 A roadmap of the dissertation

This section has been aimed at showing that reflexives, though frequently locally-bound, sometimes grammatically behave as though they take long-distance antecedents. In particular, they do so when these antecedents act as the perspective holder of the utterance, suggesting a logophoric source of non-local reference. While these interpretations have largely been associated with par-

[31] In this, I am complicit. Let my shame be known.

ticular syntactic positions and prosodic contours (Pollard & Sag, 1992; Reinhart & Reuland, 1993; Ahn, 2015), I follow Charnavel and Sportiche (2016) in suggesting that there is no principled structural division between local, and non-local interpretations of reflexive pronouns.

In light of this suggestion, we have a new way of interpreting the findings of (Parker & Phillips, 2017). While it may be the case that attention to lures is at least partially a function of morphosyntactic match (both with the lure, and with the target), there may also be a logophoric source for their findings. Notably, this would constitute a reinterpretation of their effects in terms of grammatical constraints, rather than as an error of the sentence processing mechanism. That is, incorporating logophoricity into our models of reflexive comprehension could allow us to maintain that Binding Theory strongly constrains antecedent identification—it simply isn't the only set of constraints active.

The remainder of the dissertation explores this hypothesis, with two primary goals: first, to show that sensitivity to lures is conditioned on the likelihood with which the lure is interpreted as the perspective holder; and second, to propose a model of reflexive comprehension which incorporates perspective and logophoricity. Chapters 2 through 4 address the first goal, presenting evidence that: (*i*) the kind of verb used to embed a reflexive pronoun impacts sensitivity to lures (Chapter 2); (*ii*) intervening perspective holders lessen attention to lures (Chapter 3); (*iii*) inanimate reflexive pronouns are categorically *insensitive* to long-distance lures (Chapter 4). Building on these findings, Chpater 5 addresses our second goal, proposing a modification of the cue-based architecture discussed above to accommodate logophoric constraints. In addition, Chapter 5 presents a critique of the explanatory power of cue-based models, and an re-examination of the facts surrounding co-argumenthood. This latter discussion ultimately agrees with Charnavel and Sportiche (2016) in suggesting that there is no true role for co-argumenthood in grammatical models of binding, and explores ways of accounting for the intuitions surrounding the co-argumenthood debate. Finally Chapter 6 broadens the scope of discussion to consider the implications of these findings for (psycho)linguistic theory.

CHAPTER 2

ATTITUDE VERBS IN REFLEXIVE PROCESSING

In this chapter, I explore the influence of attitude predicates on the interpretation of embedded reflexive pronouns in English. In particular, I ask whether the quality of the attitude verb affects the likelihood with which comprehenders entertain non-local interpretations of reflexives embedded in their complements. As we will see shortly, cross-linguistic facts about the distribution of logophoric pronouns might lead us to expect that non-local interpretations should be more likely when a reflexive is embedded in the complement of a speech verb (e.g. *say*, *claim*, etc.) rather than a perception verb (e.g. *hear*, *see*).

In investigating this prediction, this chapter presents two experiments, consisting of two eye-tracking while reading studies, and three off-line acceptability and interpretation surveys. Experiment 1 tests the prediction that non-local interpretations are more common when a reflexive is embedded under a speech predicate, and provides evidence that: (*i*) feature-matched lures slightly improve the acceptability of target-mismatched reflexives, but only when the embedding verb is a speech verb (Experiment 1a); (*ii*) the same qualitative pattern arises in on-line measures of processing difficulty obtained with eye-tracking while reading (Experiment 1b); (*iii*) these effects are not merely the result of a morphosyntactic feature checking operation (Experiment 1c).

Having established the effect of attitude verb-type on reflexive interpretation processes, Experiment 2 more closely examines the source of the effects observed in Experiment 1, asking whether the effect of verb-type survives when the target antecedent is not a potential perspective center. In brief, Experiment 2 finds that: (*i*) when the lure is the only potential perspective in the utterance, it uniformly impacts the acceptability of embedded reflexives, regardless of verb type (Experiment 2a); (*ii*) this same qualitative pattern obtains in eye-tracking while reading (Experiment 2b). In light of these results, I will argue that the results of Experiments 1 are not the product of intrinsic grammatical functions associated with particular (classes of) lexical items, but rather arise from the manner in which various verbs impact the assignment of a perspective center for an utterance.

The remainder of the chapter is organized as follows. Section 2.1 gives an overview of the relationship of attitude verbs to logophoricity cross-linguistically. This section includes a discus-

sion of restrictions on which verbs embed logophoric reference, and various proposals for their representation. Section 2.2 presents the results of Experiment 1, along with a brief discussion of its implications. Section 2.3 picks up this discussion and presents the results of Experiment 2. Finally, the chapter concludes in 2.4, which seeks to integrate the findings of Experiments 1 and 2 to provide an account of how attitude verbs impact reflexive interpretations in English, specifically. A more general discussion of attitude verbs and logophoric/reflexive pronouns cross-linguistically is left for Chapter 5.

2.1 Attitude verbs and logophoricity

As previously noted, logophors are pronouns which obligatorily refer to the individual whose speech, thoughts, or feelings are reported in an utterance (Clements, 1975). As a result, they are necessarily embedded under "attitude predicates": those predicates which embed a proposition the content of which is interpreted relative to the belief state of their subject, rather than the speaker (Anand, 2006; Pearson, 2015). For example, if I were to utter the sentence in (76), I would not be committed to the proposition that it is going to rain this afternoon. Instead, I'm only committed to the belief that *John* thinks this statement to be true.

(76) John thinks that it's going to rain this afternoon.

Keeping this in mind, we can slightly refine our working definition of logophors, such that the morphologically distinct logophoric pronouns of languages like Ewe, Abe, and Tupuri refer to the attitude holder of attitude predicates under which they are embedded[1] (Culy, 1994, 1997; Pearson, 2015). This definition might then be extended to include languages in which logophoricity and reflexivity are not morphologically distinguished, as in Mandarin, Japanese, Icelandic, and, I will argue, English (Huang & Liu, 2001; Kuno, 1972; Sells, 1987; Pollard & Sag, 1992; Reinhart & Reuland, 1993). Here the picture is complicated by the fact that an apparently reflexive form sometimes seems to be constrained by something akin to Principle A, but at others appears to refer to antecedents outside the standard binding domain. To treat these exceptions as instances of logophoric pronouns is thus to make the claim that they should be referring to attitude holders in the same way as morphologically distinct logophors[2].

[1]In fact the picture may be slightly more complicated than this. In Ewe, at least, a multiply embedded logophoric pronoun may refer to any of the attitude holders preceding it. Moreover, in at least some cases, it may occur in root clauses, in which case it is either interpreted as referring to the speaker, or a prominent perspective center from the previous utterance (Pearson, 2015). Both of these complications will be taken up in greater detail in Chapter 3.

[2]This picture appears to be largely true, though again not without complications. Notably, previous accounts of logophoricity in English have posited that this behavior is constrained to specific syntactic (i.e. non-argument) positions

While it appears that logophors must refer to attitude holders, languages differ with respect to which *kinds* of attitude holder are allowed to antecede a logophoric pronoun. In the first attempt at describing this variability, Sells (1987) noted that there the cross-linguistic variation in which kinds of attitude holders could anteceding a logophoric pronoun was systematic, following the implicational hierarchy in (77).

(77) SOURCE ≫ SELF ≫ PIVOT

These roles refer to various kinds of discourse entity, being the individual who makes a particular report (SOURCE), the individual whose beliefs are reported (SELF), or the individual from whose literal, physical perspective a proposition is reported (PIVOT). Critically, it is possible for all three of these roles to locate the same individual, and the individual they indicate can be either internal, or external to the sentence. Sells refines the representation of these roles by stating that they are implicationally related, such that if the SOURCE role is identified sentence-internally, the SELF and PIVOT roles must also be sentence-internal and refer to the same individual (and so on).

Using this hierarchy, Sells suggests that logophoric pronouns must refer one of these discourse roles, and that whichever role they refer to must achieve reference sentence internally. Languages then vary with respect to which particular role a logophoric pronoun attends to. A language may be SOURCE oriented, in which case all three discourse roles point to the same, sentence internal referent, or it may be SELF oriented, in which case the reference of a logophor my correspond to *either* a SELF antecedent, *or* a SOURCE antecedent (in which case, SELF and PIVOT are entailed). This has the consequence that every logophor should be capable of referring to a SOURCE discourse referent, since any lower-level orientation will be entailed by reference to SOURCE. Contrariwise, if a language allows its logophor to refer to to the PIVOT alone (to the exclusion of the other discourse roles), it must also allow that logophor to refer to either the SOURCE or SELF.

It is important to note that, in this system, the discourse roles associated with attitude holders are tied to the semantics of predicates. That is, some predicates (what Sells labels "logophoric verbs", e.g. *say*, *claim*), assign their subjects to all three discourse roles. In contrast, other predicates (e.g psych predicates like *fear* or *hope*, etc.) assign their subjects only to the SELF and PIVOT roles. Consequently, the referential potential of logophoric pronouns is (at least to some degree) isomorphic with the predicates which can embed them in a given language. If a language only

(Pollard & Sag, 1992; Reinhart & Reuland, 1993), unlike the true logophors of languages like Ewe. Similarly, the Mandarin reflexive *ziji* can refer not only to non-local attitude holders on the clausal spine, but also to an attitude-holding subject embedded *inside* an inanimate subject (Huang & Liu, 2001). To my knowledge, this configuration (known as "sub-command") is not surface-consistent with the behavior of morphologically distinct logophors in West African languages. Both of these issues will be revisited in Chapter 5 in greater detail

allows logophors to be interpreted with the SOURCE of an utterance, then that same language requires that logophors be embedded under a "logophoric" predicate. In contrast, languages which permit logophoric reference to the utterance's SELF are more relaxed, and may allow logophoric pronouns to appear beneath both logophoric and psych predicates. The exception to this is the PIVOT role, which may be assigned "constructionally" (i.e. to a particular syntactic structure), independent of a particular lexical verb.

In this respect, PIVOT (and related notions[3])has a fairly underspecified status in current models of logophoricity. First, it is the only one of Sells' roles which isn't lexically assigned, and, as Sells notes, is "free to be defined anywhere that is appropriate". Thus, there is some vaguery around when, and how the referent of PIVOT is identified. Moreover, the intended *meaning* of PIVOT seems to be more variable than other the other roles. While Sells states that it refers to the "one with respect to whose (space-time) location the content of the proposition is evaluated", it isn't clear that referents he identifies as PIVOTs always fulfill this role. For example, Sells identifies the antecedent of *zibun* (i.e. *Taroo*) as the PIVOT of the sentence in (78). Under his theory, this must be so since there is no lexical verb assigning either SELF or SOURCE, leaving PIVOT as the only possible discourse role capable of anteceding the distal *zibun*. However, *Taroo* in this sentences does not clearly represent the "physical space-time" perspective of the utterance, despite his putative role as the PIVOT anteceding *zibun*.

(78) Taroo$_i$ wa baka no Yosiko ga mizu o zibun no ue ni kobosita node
Taroo$_i$ TOP fool GEN Yosiko SUBJ water OBJ SELF$_i$ GEN on LOC spilled because
nurete-simatta
wet-got
Taroo$_i$ got wet because that fool Yosiko spilled water on him$_i$.

This problem might be solved by appealing to a slightly different set of discourse roles. For example, Charnavel and Zlogar (2015) suggest a variation on Sells' roles: ATTITUDE≫EMPATHY≫DEIXIS. This hierarchy collapses Sells' SOURCE and SELF into a single discourse role (the "attitude holder"), and expands PIVOT into two: the empathy locus, and a deictic locus. This partially solves the problem identified above—we might reasonably suppose that Taroo's emotional state is relevant for the evaluation of (78)—but doesn't directly address the question of how these discourse roles find their referents. In general, this seems to be a lacuna in theories of logophoricity: how are the various kinds of perspective centers capable of anteceding logophors established? Unfortunately,

[3]See Charnavel and Zlogar (2015)

this question lies outside the purview of the present work, and so will not be addressed here. As a result, I will continue to make use of Sells' PIVOT role, while acknowledging the outstanding issues which remain with this analysis.

2.1.1 True vs. Mixed Logophors

This property of Sells' system then aligns very neatly with an independently proposed cross-linguistic hierarchy found in (Culy, 1994). In a survey of West African languages with logophoric morphology, Culy reports a very similar implicational hierarchy in terms of which attitude predicates license logophoricity (reproduced here in 79).

(79) *speech* $\ll$ *thought* $\ll$ *knowledge* $\ll$ *direct perception*

At a glance, the analogy to (Sells, 1987) is clear, with a reasonably straightforward mapping between the verb classes identified by Culy and the discourse roles proposed by Sells. Culy's hierarchy differentiates among different varieties of SELF, breaking this role down into *thought* and *knowledge* predicates, but otherwise neatly maps speech verbs to Sells' SOURCE role, and predicates of direct perception to Sells' PIVOT. Furthermore, like Sells, Culy sees these distinctions as being semantic, rather than syntactic. He notes that expressions like *bu tame* (literally, "to bow one's head") in Ewe may be used idiomatically to mean "to think", at which point they are capable of licensing logophoricity despite lacking the literal lexical verb meaning *think*.

Despite this surface similarity, Culy maintains that the two hierarchies are not compatible, arguing that his hierarchy is necessary for what he calls "true logophoricity" (languages with morphologically realized logophoricity), while Sells' hierarchy is more appropriate for "mixed logophor" languages (languages in which reflexivity/logophoricity are encoded with the same morpheme). While he acknowledges the similarity between the two kinds of languages, he points to inconsistencies in the behaviors of Ewe and Japanese as evidence of the need to differentiate them. I recapitulate this argument in brief here.

First, he argues that the logophoric pronoun *yè* in Ewe must be PIVOT oriented, since it can occur inside causal clauses, as in (80). Importantly, causal clauses lack an attitude verb, meaning that whatever discourse role *yè* is referring to in this example must have been assigned constructionally. Recall, then, that the only one of Sells' roles capable of constructional assignment is PIVOT. Following this logic, then, *yè* in (80) must be referring to the PIVOT (to the exclusion of SOURCE and SELF), making Ewe a PIVOT oriented language. It was on the basis of similar examples (see (81)) that Sells had argued in favor of Japanese *zibun* behaving like a PIVOT oriented logophor.

(80) Kofi dzo ela bena Ama kpo yè

Kofi left because COMP Ama saw LOG

Kofi left because Ama saw him.

(81) Taroo$_i$ wa baka no Yosiko ga mizu o zibun no ue ni kobosita node

Taroo$_i$ TOP fool GEN Yosiko SUBJ water OBJ SELF$_i$ GEN on LOC spilled because

nurete-simatta

wet-got

Taroo$_i$ got wet because that fool Yosiko spilled water on him$_i$.

All else being equal, this should lead us to expect that the distribution of *yè* in Ewe, and *zibun* in Japanese to be identical. However, Culy notes that while *zibun* can occur under predicates of direct perception, *yè* cannot, as seen in (82). Assuming Culy and Sells are correct in labeling Ewe and Japanese PIVOT oriented languages, then, this pair of sentences neatly demonstrates the impossibility of cleanly mapping the discourse role PIVOT onto Culy's predicate of direct perception category. At the very least, something more would need to be said to explain why *zibun* can appear under the predicate *hear*, while *yè* cannot[4].

(82) a. Taro$_i$ wa Keiko ga zibun$_i$ no imato to hanashi o siteiru no o kita

Taro$_i$ TOP Keiko SB SELF$_i$ GEN sister DAT talk OB talking NOM OB heard

Taro$_i$ heard Keiko talking to his$_i$ sister.

b. $*$ Kofi$_i$ se koku wo le yè$_i$ dzu-m

Kofi$_i$ hear koku PRO be LOG$_i$ insult-PROG

Kofi$_i$ heard Koku insulting him$_i$.

Furthermore, Culy is concerned that distribution of mixed and pure logophors is already sufficiently distinct to warrant differential treatment. In particular, he points out that "mixed logophors" will always have a wider distribution than pure logophors, since they lead a double life as reflexive pronouns. This is a reasonable concern, and particularly pointed for this dissertation, whose central thesis is that non-clause bound behavior does in fact have its roots in true logophoricity. Unfortunately this question will continue to vex us for much of the remainder of this dissertation. As evidence accumulates that English reflexives look much more like those of

[4]It's not clear to me that more modern approaches to perspective shifting in embedded contexts couldn't handle this discrepancy. In particular, if we follow Anand (2006) and Charnavel and Sportiche (2016) in supposing that embedded operators do the work of binding logophoric pronouns, we might reasonably stipulate that the Japanese *kikoeru* embeds the necessary operator, but that *se* in Ewe does not. Even so, the point is made that collapsing Sells' and Culy's hierarchies may not be possible.

Japanese and Mandarin than previously believed, we will be continuously presented with the question of whether this class of reflexive is, in general, related to the true logophors of West African Languages. However, addressing this question immediately would, I think, detract from much of the evidence and discussion still to come (after all, we have yet to see compelling evidence of mixed-logophor behavior in English!), and so I will set it aside for the final chapter of the dissertation. For the present, it is sufficient to note that logophors (both mixed, and true) seem to be sensitive to the semantics of the predicates which embed them, and that they share a preference for referring to SOURCES (the subjects of speech verbs) over perceivers (the subjects of direct perception predicates).

2.1.2 Preliminary Evidence of Logophoricity in English

Turning now to English, we have already seen at least some preliminary evidence that English reflexives display a similar kind of discourse sensitivity. Pollard and Sag (1992), for example, note the contrast in (83), which they take as evidence that exempt reflexives are more natural when the discourse is presented from the intended antecedent's point of view, as in (83a). In terms of Sells' hierarchy, this would likely mean that non-argument anaphors in English are PIVOT oriented, since no attitude verb introduces a SOURCE or SELF in these examples.

(83) a. John was going to get even with Mary. That picture of himself in the newspaper would really annoy her, as would the other stunts he had planned.

b. ? Mary was quite taken aback by the publicity that John was receiving. That picture of himself in the newspaper would really annoy her, as would the other stunts he had planned.

Similarly, and more directly tied to the question of attitude verbs specifically, Kaiser, Runner, Sussman, and Tanenhaus (2009) present experimental evidence that non-argument reflexives preferentially refer to sources of information, rather than information recipients. In a visual-world eyetracking study, they find that participants are more likely to attend to the subject when it was source of information, rather than the recipient of information, as in (84a) relative to (84b). Put differently, comprehenders in this study preferentially attended to the subjects of speech verbs, over those of perception verbs, when assigning an interpretation to a reflexive. Again, this finding could be made to work with either of the theories discussed above.

(84) a. Peter told Andrew about the picture of himself on the wall.

b. Peter heard from Andrew about the picture of himself on the wall.

Taken together, these results in particular speak to the kinds of contrasts noted by Sells and Culy, as speakers seem to preferentially attend to attitude holders which are the subjects of speech verbs (i.e. SOURCES) over those which are the subjects of perception verbs (and thus, perhaps, PIVOTS). However, there are notable differences between these findings and the effects described in the previous section. First, the effects found in English do not appear to be as categorical as they are described for true (or truly mixed) logophor languages, where the reported judgments indicate a categorical dispreference for logophoric pronouns in unsupported environments[5]. Second, Pollard and Sag (1992) and Kaiser et al. (2009) are both concerned with reflexive pronouns in non-argument positions, a position they argue is not subject to the standard restrictions on locality for reflexives. Given this, if sensitivity to discourse factors were relegated to specific environments, this would constitute yet another difference between the facts for English and those for Mandarin, Japanese, Ewe, or other logophoric languages. Related to this latter point, finding discourse sensitivity for non-argument reflexives would not address recent findings in the sentence processing literature, where comprehenders have been found to attend to non-local referents when interpreting direct-object reflexives. In brief, current observations of discourse sensitivity in the literature, while clearly similar to the observed facts for logophors, do not find a perfect analogy there, nor do they explain other cases of Principle A incompatible behavior in on-line comprehension.

That said, there is at least suggestive evidence from the sentence processing literature that these same discourse properties are active for direct object reflexives as well. A closer examination of previous studies shows a strikingly systematic distribution in the kinds of verbs used to embed reflexive reference. As seen in Table 2.1[6], studies which failed to find evidence that comprehenders attended to lures when reading reflexives predominately embedded reflexive pronouns beneath belief (e.g. *know*, *remember*, *think*) or perception (e.g. *see*, *hear*), predicates. In contrast, the strongest evidence of lure-match facilitation to date comes from the studies conducted by Parker and Phillips (2017), in which reflexive pronouns were mostly embedded beneath predicates of speech (e.g. *say*, *claim*, *mention*). This pattern suggests two things: first, that variation in embed-

[5]This may be an artifact of judgment collecting and reporting. In which case, the hierarchies reported by Culy and Sells may be less rigid within a language than it appears from the literature. However, in the absence evidence of gradience, I will assume categorical behavior as the default.

[6]This list is not exhaustive of previous studies on reflexive pronouns, however, I believe it to be an exhaustive list of studies in which the lure was the subject of an attitude predicate embedding the target reflexive. Other studies of reflexive comprehension have embedded lures inside relative clauses modifying the main clause subject (e.g. Dillon et al., 2013; Xiang et al., 2009; Patil et al., 2016, i.a.), a position which should not be the target of logophoric reference for structural reasons. One study which had the necessary logophoric configuration but which is not included in Table 2.1 is Badecker and Straub (2002), who did not make their materials available in their paper. Notably, these authors did find evidence of lure-match facilitation, and so we would hope that, like Parker and Phillips (2017), their reflexives were predominately embedded under speech predicates.

ding verbs may be at least partially responsible for variability in the reflexives literature; second, that lure-match facilitation effects may have a logophoric source. Certainly, there is a striking resemblance of the pattern presented here to Culy's hierarchy, calling for a more direct test of the impact of attitude verbs on reflexive comprehension.

Table 2.1. Proportion verbs used to embed reflexives across studies together with whether each experiment produced lure-match facilitation

		Matrix Embedding Verb			Lure Match
Paper	**Expt**	*speech*	*belief*	*perception*	**Facilitation**
Sturt (2003b)	1	30%	54%	16%	no
Cunnings and Sturt (2014)	1	—	88%	12%	no
	2	—	86%	14%	no
Parker and Phillips (2017)	1	86%	14%	—	yes
	2	72%	28%	—	yes

2.1.3 The Logophlexives Hypothesis

The observed similarity between variability in past studies and Culy's hierarchy leads to the core hypothesis of this thesis, and to the first set of studies aimed at addressing it. Simply put, the claim is the following:

> **The Logophlexives Hypothesis:** Comprehenders will attend to lures when interpreting a reflexive pronoun if those lures can act as good logophoric antecedents for the reflexive.

The term "logophlexives" here is coined purely to remain agnostic about whether this behavior represents true logophoricity (respecting, for the present, Culy's concerns). The patterns for English found in this dissertation, like those observed previously in Mandarin, Japanese, and Icelandic, may be similar to the patterns of true logophoricity, and yet distinct from it. That said, the similarity seems striking enough to warrant in-between status of "logophlexive", at least until we tackle the matter head on at the end of this document. Consequently, the dissertation will first aim to show that English reflexives are behaving "logophor like" before turning in the end to address the question of whether this behavior represents true logophoricity.

In service of this endeavor, Experiment 1 investigates the role played by embedding verbs in controlling lure-match facilitation. At present, no study has directly assessed the impact of attitude verbs on reflexive comprehension. Experiment 1 does so by targeting the two ends of Culy's hierarchy: speech, and perception predicates. If attention to lures is controlled by logophoric principles, we should see substantially more lure-match facilitation when the reflexive is embedded under a speech predicate, than when it is embedded under a perception predicate.

2.2 Experiment 1: Attitude verbs in reflexive comprehension

To examine the effect of attitude verbs on sensitivity to principle A incompatible referents, sentences were manipulated as in (85). All items were bi-clausal, with a reflexive in the embedded direct object position. The embedded (TARGET) and matrix (LURE) subjects were then independently manipulated, so that each either matched, or mismatched the morphosyntactic features of the embedded reflexive (TARGET/LURE: $\pm$*match*). In the TARGET manipulation, mismatch was realized as disagreement with the reflexive in both number and gender, analogous to the conditions in which Parker and Phillips (2017) observed facilitatory interference. Mismatch in the LURE manipulation was realized as disagreement with the reflexive in gender. Across both manipulations, gender mismatch was accomplished with approximately half stereotypical gender (e.g. *librarian*, *janitor*), and half definitional gender (e.g. *schoolgirl*, *prince*). Finally, the matrix verb was manipulated so that it was either a verb of communication (e.g. *say*), or a verb of perception (e.g. *hear*), making the lure either a *speaker*, or a *perceiver* (VERB: *speech* vs. *perception*).

(85) The $\left\{\begin{smallmatrix}\text{librarian}\\ \text{janitor}\end{smallmatrix}\right\}$| $\left\{\begin{smallmatrix}\text{said}\\ \text{heard}\end{smallmatrix}\right\}$ that| the $\left\{\begin{smallmatrix}\text{schoolgirl}\\ \text{schoolboys}\end{smallmatrix}\right\}$| misrepresen|ted herself| at the meeting|....

If sensitivity to lure referents is conditioned on the availability of a logophoric interpretation, we should see a substantially greater lure match effect in *speech* verb sentences compared to *perception* verb sentences. For perception verb conditions, we expect a simple effect of target match, such that reflexives are read more slowly, and given lower acceptability ratings, when they mismatch the target antecedent. In contrast, lures in speech verb conditions might exert two kinds of influence on reading times at the embedded reflexive. First, a feature matched lure might *increase* reading times when the target antecedent also matches the reflexive. This would be analogous to the multiple match effect reported by Chen et al. (2012) and Badecker and Straub (2002). Second, a matched lure might *decrease* reading times when the target antecedent mismatches the reflexive, ameliorating the target mismatch penalty. This would be a replication of the finding in Parker and Phillips (2017), and analogous to the canonical findings for agreement attraction (Wagers et al., 2009). We might also expect these effects to carry over into offline measures, such that lure match has an inverse effect on sentences which are normatively grammatical (*target match*) as compared to those which are normatively ungrammatical (*target mismatch*).

It is important to note that reflexives in this experiment were always direct objects of the embedded predicates. Thus, sensitivity to lures in this study cannot be explained away as yet another instance of "exempt anaphora" in the sense of Pollard and Sag (1992) or Reinhart and Reuland (1993). Those authors posit that Binding Theory exempt behavior should only occur when an

anaphor is not co-argument with the predicates subject, which does not apply to these materials. Therefore, finding evidence of an effect of verb on lure-match facilitation would be a novel demonstration of discourse factors impacting reflexive pronouns in English, as previous demonstrations of discourse sensitivity were confined to non-argument reflexives. Finding an effect of verb type in this study would therefore strengthen the similarity between English reflexives, and logophors cross-linguistically.

To test these predictions, three separate studies were conducted using the same set of materials patterned on (85): one acceptability judgment survey, one eye-tracking while reading study, and one off-line interpretation survey. I present the results of the acceptability judgment study first.

2.2.1 Experiment 1a: Acceptability judgments

79 self-reporting native English speakers were recruited via Amazon Mechanical Turk and compensated $2 for their participation. Prior to analysis, seven participants were excluded for reporting exposure to an East-Asian language (two participants), or on the basis of age (five participants older than 55). In addition, four participants participated twice. The second instances of these subjects' participation were excluded from analysis. The remaining 68 participants were between the ages of 18 and 54 (median age: 29). 33 of these participants identified as male, and 35 as female. Participants were more or less equally distributed across 26 different states, with the exception of California, which was the home-state of 11 of participants.

2.2.1.1 Materials

48 items patterned on (85) were created and interleaved with 52 sentences from unrelated experiments in a Latin square design. In total, 24% of the items in the experiment were ungrammatical.

2.2.1.2 Procedure

The experiment was coded and hosted online using the Ibex Farm[7] software for web-based experiments. Participants were instructed to rate sentences on a scale from 1 (*very unnatural*) to 7 (*very natural*), and given four sentences exemplifying the end points of the scale (two each) as practice. Sentences were presented above the scale, with the endpoints labeled "completely unnatural" and "completely natural". There was no limit on responses. Participants indicated

[7]http://www.spellout.net/ibexfarm

their rating by either clicking the on-screen number, or pressing the corresponding number key. The experiment lasted approximately 45 minutes.

2.2.1.3 Analysis

Sentence ratings were analyzed with linear mixed effects regression, taking the factors LURE (+*match*=1, −*match*=-1), TARGET (+*match*=1, −*match*=-1), VERB (*speech*=1 vs. *perception*=-1), and all interactions as fixed effects. Random slopes and intercepts were estimated for both subjects and items, though correlations between the random effects were excluded from the model. Planned pairwise comparisons were evaluated by nesting the factor LURE inside the factors TARGET and VERB, testing for an effect of lure match within each target-match/verb-type pair. t-values of absolute value ≥ 2 were taken to be significant (Gelman & Hill, 2007).

To account for inordinately long (or short) response latencies (indicating either lack of attention, or accidental button presses, respectively) response times were z-score transformed by subject prior to analysis. Trials with $|z| > 3$ were then rejected, resulting in the exclusion of 2.2% of the data.

2.2.1.4 Results

By-subject mean ratings are given in Table 2.2. Mixed effects modeling revealed a substantial main effect of target match, reflecting the fact that participants rated target mismatch sentences significantly worse than their target match counterparts ($\hat{\beta}$=0.83, t=10.23). This main effect was qualified by a significant TARGET×LURE interaction ($\hat{\beta}$=−0.05, t=2.15). This effect was likely driven by the fact that the match-status of the lure referent had small, opposite effects on the ratings of target match, and mismatch conditions. In addition, there was a trending LURE×VERB interaction which failed to reach significance ($\hat{\beta}$=0.04, t=1.89). This trend was likely driven by the fact that feature matched lures tended to improve the ratings of reflexives in speech verb sentences overall, while having an opposite effect for target-matched reflexives in perception verb sentences. A full table of fixed effects is given in Table A.1 in the appendix. Pairwise comparisons of the lure match effect (match vs mismatch) revealed that target mismatch sentences with a speech verb were rated slightly better when the lure matched the reflexive (3.69 vs. 3.42; $\hat{\beta}$=0.14, t=2.73). No other pairwise comparisons approached significance.

Table 2.2. Experiment 1a: Mean by-subject naturalness ratings (standard error in parentheses)

		Verb Type	
Target	**Lure**	*Speech*	*Perception*
+match	*+match*	5.13 (0.11)	5.17 (0.11)
	−match	5.11 (0.14)	5.29 (0.09)
−match	*+match*	3.69 (0.14)	3.51 (0.15)
	−match	3.42 (0.15)	3.49 (0.15)

2.2.1.5 Summary

Experiment 1a revealed robust effects of target match in off-line judgments. In addition, there was no main effect of verb type, suggesting that the perception verb and speech verb sentences were overall equally natural. To the extent that we observed any influence of lure match in the offline rating task, it was confined to speech-verb conditions. However, this effect was quite small, and we failed to observe a significant VERB×TARGET×LURE interaction (despite a trending VERB×LURE interaction). Thus we cannot draw any firm conclusion about whether sensitivity to lures is modulated by verb-type in off-line judgment measures. However, it is possible that this effect is simply not durable enough to survive in off-line measures like sentence rating. Given speakers' strong dispreference for long-distance uses of reflexives in this study, it may be that effects of logophoricity are confined to a time period prior to the point at which participants are making a conscious judgment. Thus in Experiment 1b, these sentences were investigated using eye-tracking while reading to provide a more sensitive real-time measure of how comprehenders process reflexives.

2.2.2 Experiment 1b: Eye-tracking while reading

37 monolingual, English-speaking UMass undergraduates participated for extra credit in introductory linguistics and psychology courses. Details of participant exclusion are given in the analysis section.

2.2.2.1 Materials

The same materials (including fillers) from the acceptability judgment task were used in the eye-tracking study. Every item was followed by a two-alternative choice comprehension question probing aspects of the sentence other than the reference of the reflexive (by-subject question accuracy mean=83%). Comprehension questions avoided targeting the reflexives' reference to avoid alerting participants to the importance of reflexives in the study, and thus (hopefully) reduce the likelihood of their adopting an unnatural reading strategy.

2.2.2.2 Procedure

Eye movements were recorded using an EYELINK 1000 system, with a sampling rate of 1000Hz. Participants read sentences with binocular viewing, but only data from one eye was recorded. The monitor was positioned 66.3cm away from the participant, and sentences were presented in 11 point Monaco font. 3 characters were subtended by each degree of visual angle. Participants were instructed to read each sentence in a natural fashion, making sure they understood the meaning of the sentence to be able to answer the comprehension questions. After instruction, participants were given a three-point calibration in the horizontal dimension[8]. Each trial was preceded by a gaze-contingent square over the first word of the sentence. When a fixation on this square was registered, the experiment automatically displayed the sentence. Participants read each sentence at their normal rate before pushing a button on a game controller to progress to the comprehension question. To correct for calibration drift, participants were re-calibrated as necessary between trials. The experiment was preceded by four practice trials, after which participants were encouraged to ask questions about the instructions for the experiment. Together with instruction and setup, the experiment lasted approximately 45 minutes.

2.2.2.3 Regions of Interest

For purposes of analysis, two regions of interest were defined. The first was the critical reflexive region. Similar to Sturt (2003b), this region was extended to include three characters to the left of the reflexive. This analytical decision was made for two reasons. First, as noted in Sturt (2003b), function words like reflexive pronouns are frequently skipped during reading. Thus, extending the reflexive region to the left captures any possible parafoveal processing of the reflexive pronoun which may have occurred on the right-edge of the prior word. In this study, the reflexive was skipped on 16% and 6% of trials before and after extension, respectively. Extending the region of analysis therefore reduced the degree of data lost in first-pass reading measures. Second, in post-experiment interviews of participants, several reported noticing manipulations of the reflexive pronoun. This suggests that participants may have become aware of (some of) the experimental manipulations, and adopted reading strategies to cope with what were, on their face, unusual if not entirely ungrammatical sentences. Tentative evidence that this may have occurred can be seen in the rate at which participants skipped the reflexive region, which increased from 14% in the first half of the experiment to 17% in the second half. Compare this with the

[8]Since all sentences fit on a single line of text, movement in the vertical dimension was not recorded

spillover region, which was only skipped in 5% and 3% of trials in the first and second half of the experiment respectively. Together this suggests a pattern whereby participants may have adopted strategies which resulted in processing of in formation about the reflexive pronoun occurring on material adjacent to the reflexive itself. Thus, in extending the critical reflexive region to the left, we obtain a potentially more robust measure of on-line reflexive comprehension as guided by task-dependent strategies.

The second region of interest consisted of post-reflexive "spillover" material. This region contained all words after the reflexive up through the first content word (e.g. *at the meeting*). Under the assumption that some reflexive processing might persist past initial reading of the reflexive itself, effects at this region would indicate relatively late-stage correlates of reflexive interpretation.

To maintain parity across experiments, this basic regioning scheme—extended reflexive region and post-criticla region—was held constant for all eye-tracking studies reported in this manuscript.

2.2.2.4 Fixation Measures Analysis

Three standard measures of reading time were analyzed: first pass time, go-past time, and total-reading times. *First pass* times are the sum of all fixations on a region before exiting that region to the left or the right. Consequently, this measure provides a very early index of processing difficulty (when the region is sufficiently short). *Go-past* times are the sum of all fixations from when a region is first fixated until it is exited to the right, including fixations on preceding regions. This measure is frequently used as an index of recovery from an error signal, such as when a participant has recognized a target-mismatched reflexive. Go-past time thus provides a relatively early measure of processing recovery and reanalysis. Finally, *total time* measures the total sum of all fixations on a region, including any re-reading of the region. This is a very late measure of processing difficulty which includes re-reading of a region after that region has been integrated into the prior context.

Prior to statistical analysis, blinks and other artifacts were automatically removed using the robodoc software for artifact rejection developed by the UMass Amherst eyelab[9]. This software was also used to exclude trials with track-loss or blinks during first pass reading of the reflexive region. There were two participants who lost more than 25% of their data to this procedure.

[9]http://blogs.umass.edu/eyelab/software/

These participants were excluded from further analysis. 4% of the remaining data was removed due to track loss or artifacts in first-pass reading time of the reflexive region. In addition to these trial rejection criteria, fixations shorter than 80ms, or longer than 1000ms were removed prior to calculating fixation duration measures, and inordinately long first pass (>2000ms) and total time (>4000ms) values were removed from the data.

The same mixed effects model structure which was used in the acceptability judgment task was fit to each of the three fixation duration dependent measures. This included all main effects and interactions of the three manipulations (VERB: *speech*=-1, *perception*=1; TARGET/LURE: +*match*=-1, −*match*=1), with random slopes and intercepts assigned to each fixed effect by subject and item. Again, correlations among the random effects were excluded from the model.

2.2.2.5 Results

By-subject means for first pass, go-past, and total-reading time at the two regions of interest are given in Table 2.3. A graphical representation of the go-past and total-time reading measures is given in Figure 2.1. A summary of the mixed effects model fit to these dependent measures is given in Table A.2 in the appendix. Descriptively, the predictions of the logophlexives hypothesis appear to have been satisfied: reflexives which mismatched their target antecedent were read more quickly if they matched the lure referent—but only in speech verb conditions.

At the reflexive region there was a main effect of target match in both go-past ($\hat{\beta}$=55, t=3.27), and total reading times ($\hat{\beta}$=66, t=5.05), indicating longer reading times for reflexives which mismatched the target antecedent. In go-past reading times, there was also a significant TARGET×LURE interaction ($\hat{\beta}$=29, t=2.07). In both measures, these effects were qualified by a significant three-way VERB×TARGET×LURE interaction (respectively: $\hat{\beta}$=-38, t=2.5; $\hat{\beta}$=-20, t=2.33). The negative coefficients of these effect indicates that lures had a larger effect on target-mismatched reflexives embedded under speech verbs than elsewhere. Nested pairwise comparisons confirmed this interpretation, finding a lure match effect only for target-mismatch reflexives embedded under speech verbs (first-pass: $\hat{\beta}$ = 34, t = 2.05; go-past: $\hat{\beta}$ = 221, t = 4.62; total times: $\hat{\beta}$ = 121, t = 3.64). Since no other pairwise comparisons reached significance, this suggests that the strong effect of lure match on target-mismatched, speech verb reflexives drove the three-way interaction.

At the spillover region, there was a significant main effect of TARGET in go-past and total reading times (respectively: $\hat{\beta}$=97, t=5.00; $\hat{\beta}$=47, t=3.52), indicating slower reading of the spillover region when the sentence was normatively ungrammatical (target mismatch). There was also a significant TARGET×LURE interaction in go-past reading at this region ($\hat{\beta}$=31, t=2.05), indicating

lure-match facilitation in target-mismatch sentences, regardless of verb type. No other effects reached significance in this region. Nested pairwise comparisons revealed a significant lure match effect in go-past reading times for target-mismatched reflexives embedded under perception verbs ($\hat{\beta}$=141, t=2.32). No other lure-match effects approached significance (all t <1.5).

Table 2.3. Experiment 1b: By subject means for first pass, go-past, and total times at the reflexive and spillover regions (standard errors given in parentheses)

			Reflexive			Spillover		
Verb	**Target**	**Lure**	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
Speech	*+match*	*+match*	329 (17)	486 (34)	545 (28)	423 (21)	577 (47)	673 (40)
		−match	335 (16)	439 (28)	520 (30)	453 (26)	522 (32)	682 (41)
	−match	*+match*	340 (15)	460 (25)	594 (38)	437 (24)	671 (44)	752 (46)
		−match	378 (17)	682 (74)	708 (46)	414 (20)	704 (63)	762 (42)
Perception	*+match*	*+match*	336 (14)	444 (31)	502 (26)	410 (16)	558 (37)	626 (35)
		−match	343 (19)	460 (34)	502 (28)	397 (15)	554 (39)	651 (36)
	−match	*+match*	344 (15)	572 (45)	639 (37)	412 (29)	726 (52)	745 (42)
		−match	324 (10)	563 (43)	649 (36)	405 (21)	885 (70)	737 (45)

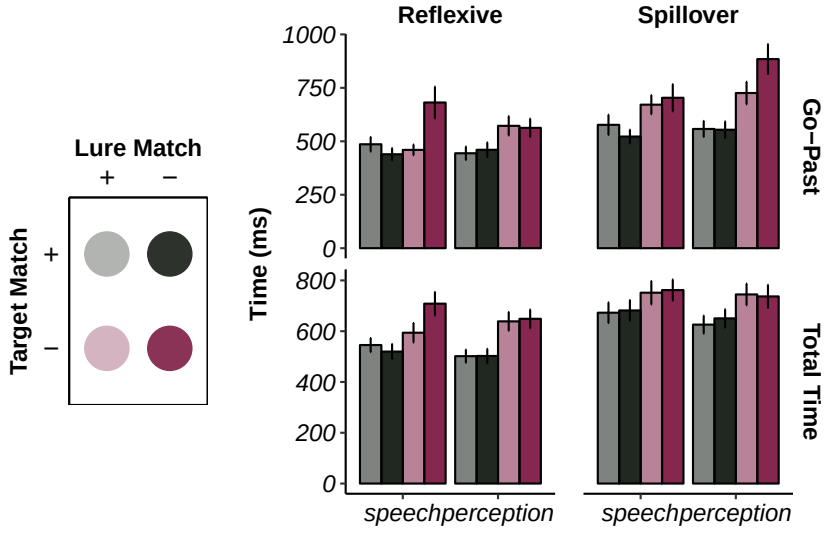


Figure 2.1. Experiment 1b: *Mean by-subject go-past and total time reading measures at the embedded reflexive region. Error bars represent standard error*

2.2.2.6 Summary of fixation duration analyses

The results of analyses performed on fixation duration measures broadly conform to the predictions of the logophlexives hypothesis. First, this experiment produced very strong lure match facilitation effects, indicating that when the local antecedent is a poor match for the reflexive, comprehenders are willing to entertain non-local antecedents. These effects were strongest at the reflexive region, where they were solely confined to sentences in which the reflexive was embed-

ded under a speech verb. This constitutes fairly decisive evidence in support of the logophlexives hypothesis: the attitude verb-type strongly constrained the availability of a non-local antecedent when comprehenders were first constructing an interpretation for the reflexive pronoun. Moreover, these attitude verbs align well with the poles of cross-linguistic hierarchy described by Culy (1994).

That said, there were at least two effects observed in these measures which do not clearly support the logophlexives hypothesis. First, the lure-match facilitation observed at the reflexive region largely disappeared at the spillover region, where very strong target-mismatch effects were observed unqualified by the manipulation of lure. While at odds with the logophlexives hypothesis, this finding actually accords well with the effects observed in Experiment 1a, where lure match facilitation, to the extent that it was present, was very small. This suggests that the interpretation assigned to that experiment, that the effect of logophoricity was insufficiently durable to survive in a judgment study, may be generally correct. This possibility, and other considerations of time course, will be addressed in Chapter 5.

The second, possibly more worrisome point is the reliable lure-match facilitation effect found for perception verb sentences in go-past reading time at the spillover. While no interaction involving VERB and LURE reached significance in this measure at this region, there was nonetheless a non-trivial pairwise facilitation effect associated with the target-mismatched, perception verb conditions. With respect to this finding, three points may be made. First, this facilitation effects is unlike the others, in that it sits on top of a fairly strong target-mismatch penalty. In the facilitation associated with speech verbs at the previous region, the lure-match, target-mismatch conditions sat more or less on par with the two grammatical baseline (target-match) conditions. In contrast, this effect seems to represent a double penalty for multiple-mismatch after a perception verb. Second, this effect is occurring at a delay relative to the effects associated with the speech verb conditions, suggesting that attention to lures is, at the very least, delayed by perception verbs. Finally, this finding may be an indication that the categorical verb preferences reported by Culy are simply more gradient in a language like English, which may not have fully grammaticized the distinctions found in languages with true logophors. This is a point I will return to in much greater detail in Chapter 5.

2.2.2.7 Cumulative progression analysis

In addition to the analysis of fixation duration measures, I examined cumulative character progression past the reflexive (Kreiner, Sturt, & Garrod, 2008; Cunnings & Sturt, 2014). Cumulative

progression is a dependent measure which indicates, at 10ms time intervals, the maximum number of characters to the right of some critical region a participant has traveled after first fixating on that region. One benefit of this measure it that it only requires the specification of a single region of interest, thereby reducing experimenter degrees of freedom. At the same time, it provides an index of processing that extends far past any initial processing of a region, and so has the power to detect effects that occur at a delay in processing, or that may be diffusely spread over a number of positions in a sentence. For our purposes, we were interested in analyzing cumultive progression data to get a clearer picture of the time course of lure match effects on sentence processing.

To analyze the cumulative progression data, the non-parametric cluster mass permutation test originally developed for ERP data by Maris and Oostenveld (2007) was implemented in R. I give here only a brief characterization of the cluster mass permutation test, and refer the reader to Maris and Oostenveld (2007) for further details. For a given pairwise comparison, this procedure identifies contiguous time points for which a two-tailed t test produces a significant t value at some arbitrary alpha level (here set to α=.9). These contiguous time points are then grouped into a "cluster". The test then determines whether a given cluster is statistically significant. To do this, t values within a cluster are summed to produce a "derived test statistic" for that cluster. The sampling distribution for this test statistic under the null hypothesis is created by randomly shuffling the condition labels on the observations and recalculating the test statistic for the cluster of interest for each permutation sample. This distribution can then be used to determine the significance of the observed test statistic for the cluster, resulting in a non-parametric test. In all simulations reported here, 1000 Monte Carlo samples were drawn to estimate a the sampling distribution of the derived test statistic under the null hypothesis.

Because we are specifically interested in *differences* between conditions, cluster mass tests were performed on difference curves that represented pairwise comparisons of continuous cumulative progression curves. This procedure was applied to the mismatch paradigm by iteratively nesting pairwise comparisons to derive the test of a three-way interaction analogous to the VERB×TARGET×LURE term in the mixed effects model. This pairwise nesting procedure is represented graphically in Figure 2.2, where each parent node represents the difference (top−bottom) of its two daughter nodes. We can interpret the derived test statistics observed at each level of comparison as follows. Positive values at the LURE comparison level represent a lure match advantage: readers have progressed further past the reflexive in lure match conditions relative to lure mismatch conditions. At the TARGET level, then, deflections from zero indicate a differential effect of LURE on target match and mismatch sentences. Finally, non-zero values at the VERB level

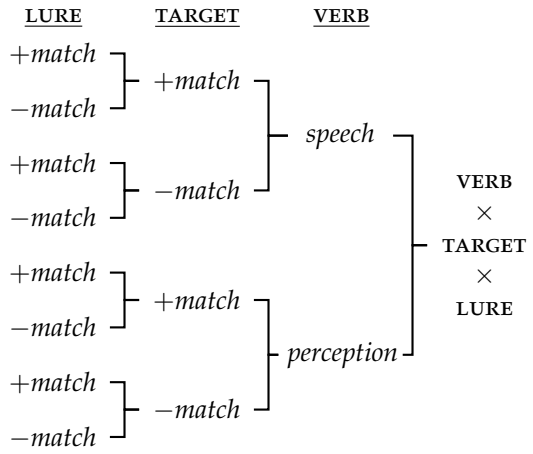


Figure 2.2. Experiment 1b: Nested pairwise contrasts calculated for the cluster mass permutation test of the three-way interaction VERB×TARGET×LURE. Each parent node is the difference (top−bottom) of its daughter nodes

indicate a difference in the TARGET×LURE interaction for speech verb conditions relative to perception verb conditions. This procedure was also carried out for the 2×2 TARGET×LURE interaction, collapsing across the VERB manipulation. All other aspects of the logic remain identical[10].

With respect to the predictions of the logophlexives hypothesis, we expect to find a significant negative derived test statistic for the TARGET×LURE interaction (a greater effect of lure match on target mismatch sentences), and a significant positive derived test statistic for the VERB×TARGET×LURE interaction (a greater TARGET×LURE interaction for *speech* verbs than for *perception* verbs).

A summary of the results of the cumulative progression analysis are given in Figure 2.3. A by-subjects cluster mass permutation test revealed a marginally significant three-way interaction of VERB×TARGET×LURE between 310ms and 430ms (p=.06). The by-items analysis found the same interaction at a slightly later time window (630–700ms; p=.05). This interaction corresponds to a larger lure-match progression advantage in this time-window for target-mismatch reflexives embedded under speech predicates than elsewhere.

In addition to this early three-way interaction, there was a substantial TARGET×LURE interaction in a later time window (by subjects: 2040–4000ms, p<.05; by items: 1900–4000ms, p<.01). This late-going two-way interaction indicates a substantial lure-match advantage for target-mismatched reflexives, irrespective of verb type. When the reflexive mismatched the target antecedent, pro-

[10] A generalized implementation of this analysis can be found at: github.com/ssloggett/cumulative_progression.

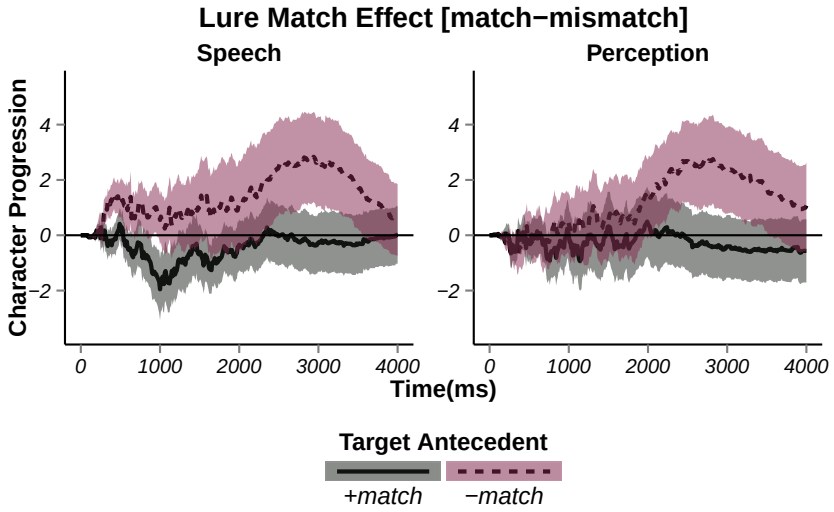


Figure 2.3. Experiment 1b: Mean by-subject lure-match effects (match-mismatch) in characters from first fixation on the reflexive. Error bars represent a bootstrapped 95% confidence interval

gression past the reflexive was eventually aided by a feature-matched lure, regardless of whether or not the matrix verb was a speech predicate. Notably, this effect occurred at a substantial delay relative to the initial three-way interaction. In this sense, it seems similar in character to the lure-match facilitation effect for target-mismatched reflexives in perception verb sentences observed earlier in go-past reading times at the spillover. Together, these effects lend credence to the notion that the effect verb-type observed in this experiment is gradient, rather than categorical.

2.2.2.8 Summary

The findings from eye-tracking while reading provide striking support for the hypothesis that a logophoric representation drives violations of Binding Theory in comprehension. In the analysis of fixation duration measures, we observed that reflexives were read more slowly when the target antecedent was a poor morphosyntactic match. Importantly, this target mismatch penalty was preferentially ameliorated by a feature-matched lure referent when the reflexive was embedded under a speech verb, an effect found in all three reading time measures. This finding was replicated in an early time window of the cumulative progression analysis, where a feature-matched *speaker* facilitated progression past a target-mismatched reflexive, but a feature-matched *perceiver* did not. These findings suggest that the lure referent is more accessible when it is the subject of a speech verb than when it is the subject of a perception verb. This selective fallibility is consistent with a logophoric analysis of the reflexive from.

However, while these data provide compelling evidence that attitude predicates impact sensitivity to lure referents, they do not speak to the question of whether comprehenders are actually entertaining logophoric interpretations of embedded reflexives. Notably, these data are consistent with two alternative explanations:

(*i*) Comprehenders use logophoric centers to check a reflexive's morphosyntactic features

(*ii*) Comprehenders actively interpret logophoric centers as the antecedents of reflexive forms

The results obtained in eye tracking are compatible with the view that comprehenders are interpreting reflexives logophorically, they do not decisively demonstrate this. It remains possible that some low-level feature checking operation (sensitive to attitude predicates for reasons unknown) is responsible for the effects observed in Experiment 1b. To address this issue, Experiment 1c adapted the materials from the previous studies for use in an off-line interpretation survey designed to probe comprehenders' interpretations of embedded reflexives.

2.2.3 Experiment 1c: Interpretation survey

64 self-reporting native English speakers were recruited via Amazon Mechanical Turk and compensated $4 for their participation. Prior to analysis, 19 participants were excluded for reporting exposure to an East-Asian language, age (older than 55), or for indicating prior participation in a study about reflexive pronouns. The remaining 45 participants were between the ages of 22 and 55 (median age: 33).

2.2.3.1 Materials

The same items (and fillers) from Experiments 1a and 1b were used in this experiment.

2.2.3.2 Procedure

Sentences were presented to participants in a "chunked" self-paced reading paradigm, wherein participants pressed a button to reveal progressive groups of words in the sentence. Sentence chunks were presented in the center of the screen, and with each button press the previous chunk of text was replaced by the subsequent chunk, until the sentence was completed. At the end of each sentence, participants were asked a binary-choice comprehension question. For filler sentences, these questions targeted various pieces of the sentence (e.g. facts about the subject, object, predicate, or thematic roles involved). For target-mismatch sentences, these questions probed aspects of the interpretation which did *not* rely on participants' interpretation of the critical

reflexive. However, for target-match conditions, the comprehension question explicitly targeted the object position of the embedded predicate, and presented the two subjects (embedded and matrix) as possible answers to the question. A sample sentence/question pair of this type is given in (86).

(86) The $\left\{\begin{array}{c}\text{librarian}\\ \text{janitor}\end{array}\right\}$ $\left\{\begin{array}{c}\text{said}\\ \text{heard}\end{array}\right\}$ that the schoolgirl misrepresented herself at the meeting...

Who was misrepresented at the meeting?

the schoolgirl *the librarian/janitor*

Critically, a participant's choice of answer to this question indicates their chosen interpretation of the embedded reflexive. In this way, the task was similar to that employed by (Sturt, 2003b). Questions probing the reflexive interpretation were only asked after target-match sentences to discourage participants from adopting a response strategy. Had participants seen questions probing their interpretation of target-mismatched reflexives, they may have been induced to artificially favor non-local interpretations. Moreover, in probing the interpretation of target-matched reflexives, we gain a measure of the base-rate at which comprehenders entertain non-local interpretations even when they are not forced to do so: providing a matrix answer to the question in (86) represents an unforced error in the part of the comprehender.

Finally, all other aspects of the manipulations employed previous remained unchanged, allowing for an evaluation of the role of verb-type in promoting non-local interpretations. As before, we predict more lure responses to comprehension questions when the lure matches the reflexive, and when the embedding verb is a speech verb.

2.2.3.3 Analysis

A logistic mixed effects model was fit to proportion matrix responses for the subset of the data corresponding to the target-match conditions. The factors LURE, VERB, and their interaction were taken as fixed effects, with uncorrelated random slopes and intercepts fit to subjects and items.

2.2.3.4 Results

The by-subject mean proportion lure responses are summarized in Table 2.4. The results of the mixed effects model are presented in Table A.1 in the appendix. This model revealed significant main effects of LURE (-0.34, $p < .001$) and VERB (-0.24, $p < .01$), such that participants were more

likely to give a lure response when it matched the reflexive, and when the embedding verb was a speech verb. However, these effects appeared to be additive, as the interaction term did not approach significance. Nested pairwise comparisons to assess the affect of lure-match within verb-type revealed a significant effect for both speech ($\hat{\beta}$=-.33, p¡.01) and perception ($\hat{\beta}$=-.33, p¡.01) verbs, indicating more non-local interpretations for both verb types when the lure matched gender of the reflexive.

Table 2.4. Experiment 1c: Mean by-subject proportion lure responses

	Verb Type	
	Speech	*Perception*
+match	0.34	0.26
-match	0.23	0.16

2.2.3.5 Summary

The results of this interpretation survey demonstrate two important findings. First, these findings replicate the primary result of Experiment 1b, demonstrating that comprehenders are more likely to consider lure referents which are the subjects of speech verbs. Second, it clarifies the source of the effects observed in Experiment 1b, suggesting that comprehenders are indeed entertaining long-distance interpretations of embedded reflexives. In this respect, these findings stand at odds with the superficially similar class of "grammatical illusions" like agreement attraction (Wagers et al., 2009, i.m.a.). In these cases, comprehenders fail to notice ungrammatical number agreement in sentences like (87). The standard explanation for this effect is that comprehenders spuriously license the plural marking on the verb (here "were") with the features of a non-subject noun (here "cabinets"). As noted elsewhere (Dillon et al., 2013), this pattern is at least descriptively similar to the lure-match facilitation effects we've been discussing. However, There is very little evidence that attraction errors occur concomitant with interpretive errors (i.e. comprehenders don't mistake the cabinets has having been described as "rusty with disuse"; Schlueter, Parker, & Lau, 2017). The results of Experiment 1c thus join a growing body of work (Dillon et al., 2013; Andrews, Yacovone, Sloggett, & Dillon, 2016) demonstrating that these two patterns of behavior are fundamentally different.

(87) The key to the cabinets definitely were rusty with disuse.

Moreover, this experiment demonstrates that comprehenders are entertaining non-local interpretations even when the local subject is a good morphosyntactic match. This provides a striking contrast to previous work which suggested that lure referents were only entertained in the pres-

ence of gross target mismatch (Parker & Phillips, 2017). The fact that reliable lure-match effects in target-matched sentences remain elusive in measures of reading difficulty will join the discussion of time course and gradience in Chapter 5.

One worrying aspect of these data is the high rate at which participants gave lure responses even when the lure didn't match the morphosyntax of the embedded reflexive (22% and 17% for speech and perception verbs, respectively). However, there are several potential explanations for this fact. First, it is important to note that the main effect of LURE was significant, indicating that participants were sensitive to the morphosyntactic features of the lure referent. Second, the feature match of the lure and reflexive was achieved with a difference in gender, a feature known to be less reliable as a disambiguating cue (Carminatti 2002). Finally, the gender manipulation of the lure was achieved with stereotypical gender in approximately half the items. For these items, the lure was only an impressionistically bad antecedent, rather than a true morphosyntactic mismatch. Given this, it is possible that comprehenders entertained this antecedent despite its surface incompatibility with the reflexive.

2.2.4 Discussion

Taken together, Experiments 1a-1c provides compelling initial evidence that reflexive comprehension is sensitive to discourse parameters very similar to those observed for logophoric pronouns elsewhere. In Experiment 1a, we saw that, to the extent that lure-match facilitation was observed, it was confined to reflexives embedded under speech verbs. In Experiment 1b, we saw evidence from four different dependent measures of reading difficulty that lure-match facilitation was significantly earlier and stronger for reflexives embedded under speech verbs than for those embedded under perception verbs. Finally, in Experiment 1c, we saw that these effects are very plausibly driven by actual interpretive considerations, rather than mere morpho-syntactic feature checking. Taken together, this is strong supporting evidence in favor of the logophlexives hypothesis advanced here.

That said, aspects of Experiment 1 remain puzzlingly at odds with some of the predictions drawn in Section 2.1. In particular, we observed three pieces of evidence that although perception verbs dampen sensitivity to lures they do not entirely remove it. First among these, we observed in Experiment 1b a significant lure-match facilitation effect for target-mismatched reflexives embedded under a perception verb. While this effect only emerged in go-past reading times at the spillover region, a similar effect was then observed in a late time-window of the cumulative progression analysis, suggesting that the effect is present, albeit, at a delay. Finally, a substantial

lure-match effect was observed with perception verb sentences in Experiment 1c. These effects do not align with the nicest[11] predictions of the logophlexives hypothesis, indicating that English speakers don't locate themselves somehwere on either of Culy's or Sells' hierarchies and stick there.

In accounting for these data, there seem to be two options. First one might suppose that, given the delay associated with perception verbs, these findings represent an altogether different kind of process from the one observed with speech verbs. Perhaps it is the case that speakers initially don't consider the subjects of perception verbs, or do so only when they are unhappy with the local subject as an antecedent. That is, faced with an unhappy locally bound reflexive, comprehenders consciously seek out an alternative, and are less choosy with their antecedents when they do so. Sturt (2003b) gave a similar explanation for the lure-match effects he observed in later reading measures. Regrettably, Experiment 1c seems to thwart this interpretation. First, the critical data from this experiment revolved around reflexives which matched their local antecedent, in which case, presumably, comprehenders shouldn't need to go searching for a better referent. However, assuming that on some percentage of trials they do so any way (boredom in an experiment being what it is), we might then expect that lure responses in perception verb conditions should be slower relative to lure responses in speech verb conditions. This was not observed[12]. In Experiment 2, I explore an alternative explanation of these findings.

2.3 Experiment 2: De-confounding perspective and thematic role

In Experiment 1 we were surprised to find that, at least some of the time, the subjects of perception verbs are considered as referents for embedded reflexives. One possible explanation for the data is that attitude verbs in English are doing something rather less categorical than they are in logophoric languages. In true instances of (mixed) logophoricity, verbs serve the role of assigning their subjects to one of a variety of perspective center (e.g. SOURCE, SELF, PIVOT), and it is this perspective center which is targeted by logophoric reference. For Culy and Sells, differences among langauges arise from differences in which kinds of perspective center a logophoric pronoun can refer to. However, another possible source of variation lies in whether verbs grammatically assign their agents to these roles or not.

[11] In the sense of Agnes Nutter's "Nice and Accurate Prophecies"

[12] Analysis pending

Sells (1987) suggests that the role PIVOT may be assigned constructionally. Indeed, he goes so far as to suggest that "PIVOT is not lexically specified, [so] it is essentially "free" to be defined anywhere that is appropriate" (Sells, 1987). Perhaps English doesn't rigidly designate assignment to any of Sells' discourse roles. Instead, it may rely on *probabilistic* information to guide its choice of perspective center. If so, it may leverage information like verb class to determine who the perspective center of an utterance is. This would be particularly important in utterances with more than one perspective utterance, where subtle cues may be necessary to determine from whose perspective a proposition is intended to be evaluated. Under this (still sketchy) version of events, the results of Experiment 1 represent a probabilistic preference for interpreting the subject of a speech verb as the perspective center of the sentence, leading to its accessibility to the embedded reflexive.

One notable aspect of Experiment 1 is that in all cases there were two technically possible perspective centers for the utterance: both the lure, and the target were animate consciousness centers, and therefore capable of holding the perspective relevant for interpreting the utterance (Charnavel, p.c.). In this situation, comprehenders may be more likely to make use of the cues made available by attitude verbs in deciding whose perspective is relevant. In contrast, If one or the other referent were inanimate, then the choice of perspective center would be clear: the animate referent, being the only sentient entity, would be forced to hold that role.

This observation allows for a testable prediction. If English, like Ewe and Japanese, has grammaticized the perspective holders to which a reflexive (or logophor) can refer, then the animacy of the target antecedent should be irrelevant. Reference to the matrix subject should be acceptable under speech verbs, and unacceptable under perception verbs. On the other hand, if attitude verbs are helping guide the assignment of perspective in the face of (in principle) ambiguity, then the presence/absence of alternative potential perspective centers should matter a great deal. In particular, if speech verbs in Experiment 1 were only helping to *signal* perspective (rather than forcibly assigning it), then removing the choice of perspective center should obviate their effect.

This is the prediction Experiment 2 is designed to test: Does the effect of verb survive when the lure is the only *possible* perspective holder in the utterance? If particular verbs grammatically license non-local reference, then it should. If attitude verbs merely serve as cues to perspective, then it shouldn't. To address these possibilities, the materials from Experiment 1 were adapted as shown in (88). First, each stimulus consisted of a context sentence, and a test sentence. The context sentence introduced a referent which served as the antecedent for the embedded subject (TARGET) of the test sentence. So, for example, if the context sentence introduced a "journalist", the target

antecedent of the test sentence was a proper name (like "Jill") which matched the gender of the embedded reflexive. If the context sentence introduced an inanimate referent (e.g. an "article"), the target antecedent of the test sentence was the inanimate pronoun "it", referring back to the referent introduced in the context sentence. Note that in these conditions, the reflexive always mismatched the target antecedent. In addition to this (somewhat baroque) manipulation, the lure referent was manipulated so that it either matched, or mismatched the reflexive in gender. As in Experiment 1, gender (mis)match was approximately half stereotypical, half definitional. Finally, in target-mismatch conditions *only*, the matrix verb was manipulated so that it was either speech verb (e.g. "say") or a perception verb (e.g. "hear"). More concisely, there were three levels of TARGET in this study (*target +match; target −match (speech); target −match (perception)*), and two levels of LURE (±*match*). Fully crossed, this yielded six conditions, allowing us to test for an effect of lure in target match sentences, and differential effects of lure in target-mismatch sentences based on verb-type.

(88) **Context:** The salacious journalist/article was widely derided.

a. The $\underline{\textbf{lure}}$ $\left\{\begin{array}{c}\text{actress}\\ \text{actor}\end{array}\right\}$ said that Jill lied about herself...

b. The $\underline{\textbf{lure}}$ $\left\{\begin{array}{c}\text{actress}\\ \text{actor}\end{array}\right\}$ $\underline{\textbf{verb}}$ $\left\{\begin{array}{c}\text{said}\\ \text{heard}\end{array}\right\}$ that it lied about herself...

Critically, in target-mismatch sentences the target was actually an inanimate referent. Since inanimate referents are incapable of being perspective centers, they cannot antecede logophors, and therefore cannot compete with the matrix subject for reference. Consequently, finding an effect of verb-type in these conditions would constitute evidence of grammaticization, while failing to find it would indicate a more gradient role for attitude verbs.

Two separate studies using materials based on (88) were conducted: one off-line acceptability judgment survey, and one eye-tracking while reading study. I present the result of the acceptability survey first.

2.3.1 Acceptability judgments

54 self-reporting native English speakers were recruited via Amazon Mechanical Turk and compensated $5 for their participation. Prior to analysis, 1 participant was excluded due to age (older than 55). All remaining participants reported no prior experience with a reflexive exper-

iment or exposure to an East Asian language. The remaining 53 participants were between the ages of 22 and 53 (median age: 32).

2.3.1.1 Materials

36 sets of sentences patterned on (88) were created and interleaved with 24 items from an unrelated experiment and an additional 48 fillers. All non-test items were grammatical, resulting in a total of 22% normatively ungrammatical sentences in the experiment. Fillers were controlled such that 24 items contained "it" as the embedded subject of an attitude predicate. For half of these (12 items), the continuation contained a grammatical use of the inanimate reflexive "itself". For the remaining half, the sentence continued with a standard, non-reflexive direct object. The remaining 24 fillers consisted of bi-clausal sentences lacking a reflexive in the embedded object position. These measures were taken to ensure reasonably equitable distribution of reflexive and inanimate pronouns in the experiment in an attempt to reduce the likelihood of task-dependent strategies. All stimuli in the experiment, including fillers, consisted of a context/test sentence pair.

2.3.1.2 Procedure

The experiment was coded and hosted on the Ibex Farm server for web-based experiments. Participants were presented pairs of sentences on the screen. The context sentence was always on top, and separated from the test sentence by two lines of white space. Both sentences were labeled "Context:" and "Target:" respectively. Participants were instructed to read both sentences carefully, and then rate the naturalness of the target sentence only on a likert scale of 1-7. The end-points of the scale were labeled "completely unnatural" and "completely natural", respectively. As in Experiment 1a, the stimulus was presented above the scale and visible when participants made their judgment. There was no time limit on responses. Prior to participation, participants completed a brief practice session consisting of four sample stimuli, used to ground the end-points of the naturalness scale. Including this practice session, average completion time was approximately 45 minutes.

2.3.1.3 Analysis

A linear mixed effects model was fit to sentence rating for the experimental items. The factors LURE, TARGET, and their interaction were used as fixed effects, with uncorrelated random slopes and intercepts fitted to subjects and items. The two-level factor LURE was sum coded

(+*match*=1, −*match*=-1) to test for an effect of lure match. The three-level factor TARGET was helmert coded with two separate contrasts, one probing for an effect of target match (TARGET: *Name* = 1; *Speech*/*Perception*=-.5), and another for an effect of verb-type (VERB: *Speech* = 1; *Perception* = -1).

2.3.1.4 Results

By-subject mean naturalness ratings are given in Table 2.5. A full table of fixed effects for the model fit to these data is presented in Table A.4, in the appendix. This model revealed significant main effects of TARGET ($\hat{\beta}$0.89, t=6.35) and LURE ($\hat{\beta}$=0.16, t=3.63), indicating increased naturalness ratings when (1) the target antecedent matched the reflexive, and (2) when the lure matched the reflexive, respectively. These main effects were qualified by a significant TARGET×LURE interaction ($\hat{\beta}$=-0.24, t=3.77), indicating a larger effect of lure on sentences containing a target-mismatched reflexive relative to those containing a target-matched reflexive. Nested pairwise comparisons evaluating the effect of lure match within the levels of TARGET found a significant lure-match advantage for target-mismatched reflexives embedded under both speech ($\hat{\beta}$=-0.45, t=3.81) and perception ($\hat{\beta}$=-0.68 t=-5.71) verbs. The lure-match effect for target-matched reflexives did not approach significance.

Table 2.5. Experiment 2a: Mean by-subject naturalness ratings (standard error in parentheses)

	Lure	
Target	*+match*	*−match*
Name	5.11 (0.16)	5.28 (0.13)
Speech	4.16 (0.15)	3.69 (0.16)
Perception	4.12 (0.16)	3.44 (0.17)

2.3.1.5 Summary

As in Experiment 1a, this experiment demonstrates a strong preference for local reference with reflexives. Participants were loathe to assign high ratings to sentences with target-mismatched reflexives. However, unlike in Experiment 1a, this experiment found a substantial effect of lure-match, demonstrating that target-mismatched reflexives can be (at least partially) rescued by a feature-matched lure, even in off-line measures of acceptability. Also at odds with Experiment 1a: this study found no moderation of the lure match effect as a function of verb-type. Feature-matched lures rescued target-mismatched reflexives embedded under speech and perception verbs alike. While there was a trending effect of VERB, such that speech verb sentences were rated better over all, this effect failed to reach significance (t=1.72). In brief, these findings lend no support

to the hypothesis that verbs grammatically encode the perspectival properties of their subjects. Rather, this is suggestive evidence in favor of the view that verbs provide cues to perspective, rather than assigning it[13]. However, as in Experiment 1a, the possibility remains that effects of logophoricity may be relegated to earlier stages of comprehension not accessible via off-line judgment surveys. So, as before, we turn to eye-tracking.

2.3.2 Eye-tracking while reading

45 monolingual, English speaking UMass undergraduates participated for (extra) credit in introductory linguistics and psychology courses. Details of participant exclusion are given in the analysis section.

2.3.2.1 Materials

The same materials (including fillers) from the acceptability judgment task were used in the eye-tracking study. Every time was followed by a two-alternative choice comprehension question. Half of these questions probed facts about the context sentence. The remainder probed aspects of the test sentences other than the reference of the reflexive (by-subject question accuracy mean = 88%). As before, comprehension questions avoided targeting reflexives to avoid drawing attention to this manipulation.

2.3.2.2 Procedure

The same EYELINK system and setup as in Experiment 1b were used in Experiment 2b, with the exception that participants were instructed that they would be reading pairs of sentences, rather than sentences in isolation. Trials always consisted of three separate screens, one for the context sentence, one for the test sentence, and a final screen containing the comprehension question. The context and test sentences were preceded by a gaze-contingent box which acted as a calibration check. All other aspects of the setup were identical to Experiment 1b.

2.3.2.3 Analysis

As in Experiment 1b, first pass, go-past, and total reading times were analyzed for the reflexive and spillover regions. The reflexive region was the embedded reflexive, extended three characters to the right to accommodate high skipping rates (20% vs. 10% vefore and after extension, respec-

[13] Although, a word of caution: this is reasoning based on a null-result, and as such needs bayesian support (or something else) to gain fuller support.

tively). The spillover region contained the material to the right of the reflexive, up through the first content word.

Trial and artifact rejection criteria were identical to those used in Experiment 1b. Six participants were excluded from analysis due to excessive data loss (more than 25%). X% of the remaining data was removed due to track loss or artifacts on the reflexive region during first-pass reading. In addition, inordinately long first pass and total time values were removed from the data (>2000ms or 4000ms, respectively).

The same mixed effects model structure used in the acceptability judgment task was adopted for each of the three fixation duration measures. This included hte main effects and interactions of LURE (+*match*=-1, −*match*=1), and the helmert coded factor TARGET (TARGET: *Name*=-1, *Speech/Perception*=.5; VERB: *Speech*=-1, *Perception*=1). Random slopes and intercepts were assigned to subjects and items, excluding correlations among the random effects.

2.3.2.4 Results

By-subject means for first pass, go-past, and total-reading time at the two regions of interest are given in Table 2.6. A graphical representation of the go-past and total-time reading measures is given in Figure 2.4. A summary of the mixed effects model fit to these dependent measures is given in Table A.5 in the appendix.

At the reflexive region, there were significant main effects of TARGET and LURE in total reading times (respectively: $\hat{\beta}$=28, t=2.16; $\hat{\beta}$=42, t=2.44). These effects were qualified by a significant TARGET×LURE interaction, also in total reading time ($\hat{\beta}$=50, t=2.17). No other effects approached significance. Nested pairwise comparisons to evaluate the effect of lure match revealed significant lure match effects for target-mismatched reflexives in both speech verb ($\hat{\beta}$=34, t=2.52) and perception verb ($\hat{\beta}$=33, t=2.45) sentences, suggesting that lure-match effects in target-mismatch sentences drove the TARGET×LURE interaction.

At the spillover region, there were again main effects of TARGET and LURE in go-past and total reading time (TARGET: $\hat{\beta}$=104, t=4.19; $\hat{\beta}$=48, t=3.06; LURE: $\hat{\beta}$=115, t=3.01; $\hat{\beta}$=47, t=2.69; respectively). As before, these main effects were qualified by a significant TARGET×LURE interaction in both measures (respectively: $\hat{\beta}$=135, t2.57; $\hat{\beta}$=91, t=3.59). Nested pairwise comparisons revealed significant effects of lure match in go-past reading times (speech: $\hat{\beta}$=105, t=3.36; perception: $\hat{\beta}$=77, t= 2.45), and in total reading times (speech: $\hat{\beta}$=60, t=4.13; perception: $\hat{\beta}$=33, t=2.26). Again, these findings suggest that indiscriminate attention to lures in target-mismatch sentences drove the TARGET×LURE interaction.

Table 2.6. Experiment 2b: By subject means for first pass, go-past, and total times at the reflexive and spillover regions (standard errors in parentheses)

		Reflexive			Spillover		
Target	**Lure**	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
Name	*+match*	301 (15)	395 (26)	413 (24)	382 (20)	490 (72)	521 (30)
	−match	293 (13)	367 (19)	405 (17)	361 (17)	462 (33)	478 (26)
Speech	*+match*	299 (12)	478 (62)	423 (21)	368 (17)	538 (37)	514 (26)
	−match	310 (15)	447 (35)	497 (30)	398 (17)	748 (79)	636 (41)
Perception	*+match*	290 (12)	404 (31)	405 (18)	391 (22)	549 (53)	532 (28)
	−match	304 (14)	397 (22)	474 (25)	407 (21)	711 (59)	599 (25)

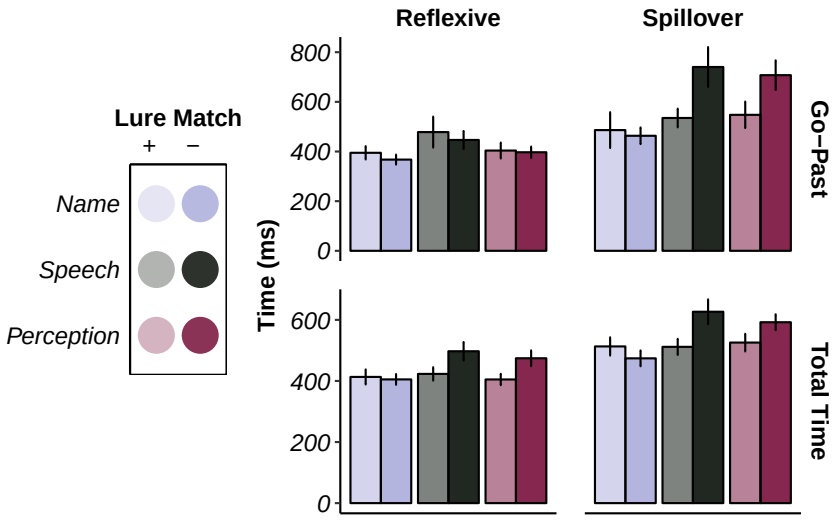


Figure 2.4. Experiment 2b: Mean by-subject go-past and total time reading measures at the embedded reflexive and spillover regions. Error bars represent standard error

2.3.2.5 Summary of fixation duration analyses

Analyses performed on fixation duration measures found substantial lure-match facilitation effects for target-mismatched reflexives regardless of verb-type. These effects emerged in both go-past and total reading time at the spillover region, and in total-reading times at the reflexive region. While these effects are later than those observed in Experiment 1b (both in terms of the measure involved, an in the locus of the effect), there is relatively little evidence indicating an error-repair strategy in these data. That is, we see no conclusive evidence that comprehenders first noticed the target-mismatch, then actively engaged in strategies to ameliorate that error signal. While there was a trending main effect of TARGET in go-past reading times at the reflexive, this effect failed to reach significance (t=1.9). In light of this, relatively little can be made (at present)

of the difference in time course between Experiments 1b and 2b. However, more space will be devoted to this discussion once the results of Experiments 3c and 4b have been presented.

In sum, these findings accord well with the results of the acceptability survey presented in Experiment 2a. We see no evidence of differentiation among kinds of attitude verb when the lure is the only possible perspective holder in the utterance. That said, it is interesting to note that the lure match effects associated with speech verb conditions ($\hat{\beta}$=34, 105, 60) were systematically larger than those associated with perception verb conditions ($\hat{\beta}$=33, 77, 33), suggesting that this study may not have been high-powered enough to detect a VERB×LURE interaction. Regardless, these effects leave little doubt that lure-match facilitation *can* be found with perception-verb subjects, suggesting that the effects observed in Experiment 1b were not categorically grammatical in nature. While speech verbs may yet promote long-distance reference, perception verbs cannot be said to categorically disallow it.

2.3.2.6 Cumulative progression analysis

As in Experiment 1b, a cumulative progression analysis of fixations past the reflexive was also conducted. Once again, the cumulative progression curves were statistically evaluated using a cluster mass permutation test, assuming an α cutoff of 0.9 for cluster identification. This test evaluated difference-of-differences equivalent to the TARGET×LURE and VERB×LURE interaction in the mixed effects model. These interaction terms were the result of the nested pairwise comparisons shown in Figure 2.5, comparing the relative effect of lure match (*match*−*mismatch*) on target match/mismatch sentences (*Name* vs. *Speech/Perception*), and on reflexive embedded under speech verbs relative to those embedded under perception verbs (*Speech* vs. *Perception*).

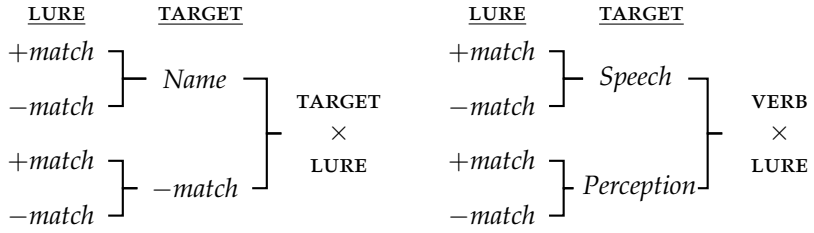


Figure 2.5. Experiment 2b: Nested pairwise comparisons calculated for separate cluster mass permutation tests of TARGET×LURE and VERB×LURE. Each parent node is the difference (top-bottom) of its daughter nodes

A graphical representation of the lure-match effect in character progression for each level of TARGET is given in Figure 2.6. The by-subjects cluster mass permutation test revealed a significant

lure-match advantage for reflexives embedded under speech verbs between 780ms and 3310ms (p <.01). Likewise, there was a significant lure match advantage in perception verb conditions between 380ms and 1400ms (p <.01), and at a later time window between 2300ms and 33600ms (p <.05). These lure-match advantages were replicated in the by-times analysis at analogous time windows (speech: 800-3900ms, p <.01; perception: 2150-3900ms, p¡.05).

Collapsing across verb types, there was an overall lure-match advantage for target-mismatched reflexives (by subjects: 210-3470ms, p <.01; by items: by-items: 360-3860ms, p < .05). The by-items analysis revealed a marginal, early lure-match advantage for target-matched reflexives (110-160ms, p=.07), but this effect was absent from the by-subjects analysis.

Turning to the interactions of interest, the by-items analysis revealed a marginal VERB × LURE interaction between 1570ms and 1970ms (p=.06), but this effect was absent from the by-subjects analysis. However, both analyses revealed a significant TARGET × LURE interaction at a comparatively late time window (by-subjects: 1000-3310ms, p <.05; by-items: 1650-3440ms, p <.05), indicating a larger lure-match effect associated with target-mismatch reflexives, relative to target-match reflexives.

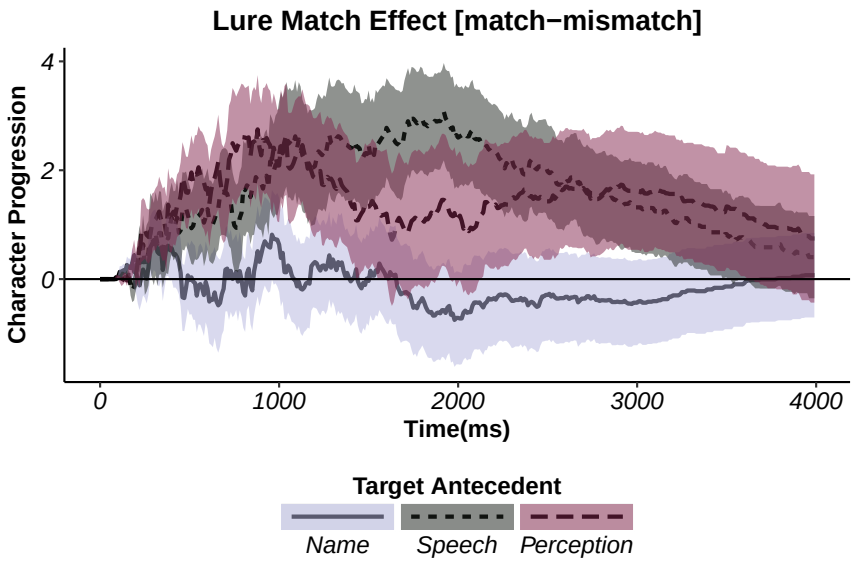


Figure 2.6. Experiment 2b: Mean by-subject lure-match effects (match-mismatch) in characters from first fixation on the reflexive. Error bars represent a bootstrapped 95% confidence interval

In sum, the results of the cumulative progression analysis largely replicate the findings of the more traditional, fixation-duration based analyses. We see strong evidence of lure-match facilitation associated with both speech and perception verb conditions. There is a hint at an

interaction of VERB×LURE, but nothing decisive enough to draw conclusions from. However, the TARGET×LURE action is reasonably robust, and similarly late. The primary generalizations of Experiment 2b thus seem fairly clear: in the absence of multiple possible perspective centers, comprehenders attend to lures regardless of verb type. This effect is surprisingly late, but does not, at present, seem to be (entirely) a repair strategy.

2.3.3 Discussion

Overall, the results of Experiment 2 are extremely consistent. In both off-line acceptability rating, and on-line measures of reading difficulty, comprehenders show a dispreference for reflexives which mismatch their target antecedent. This dispreference is ameliorated by the presence of a feature-matched lure, irrespective of verb-type. It is interesting to note that while lure-match facilitation for reflexives embedded beneath perception verbs was observed in Experiment 1b, it was at a delay, and failed to *completely* ameliorate the violation. In contrast, lure-match in Experiment 2b produced amelioration that rendered reading times more or less on-par with the grammatical baselines. Together with the results of Experiment 2a, this suggests that long-distance reference embedded under perception verbs isn't categorically disallowed. Instead, it seems to be the case that attitude verbs can influence the likelihood with which comprehenders entertain long-distance interpretations when more than perspective holder available.

2.4 Incorporating attitude verbs into processing models

With the results of Experiments 1 and 2 in hand, we can begin to draw slightly firmer conclusions about the role of logophricity in reflexive processing. First, there is now compelling evidence that attention to lure referents is not a uniform function of degree of target-mismatch (c.f. (Parker & Phillips, 2017)). Comprehenders are more likely to consider lure referents which are the subjects of speech verbs (Experiments 1a, 1b) even when the target antecedent is a perfect morphosyntactic match (Experiment 1c). This strong effect of verb type disappears when the target antecedent is inanimate, leaving the lure referent as the only possible perspective holder in the utterance (Experiments 2a, 2b). From the point of view of perspective taking, these results tell a compelling, and consistent story: comprehenders attend to lures which are the (preferred) perspective center of an utterance. This observation resonates strongly with the central generalization about logophoricity: logophoric pronouns refer to the entity whose speech, thoughts, or beliefs (i.e. perspective) are represented in an utterance. Given this, Experiments 1 and 2 provide fairly strong evidence in favor of the logophlexives hypothesis, as laid out in section 2.1. Comprehenders attend to lures

which make good anchors for perspective, in the same way that logophoric pronouns appear to pick out the attitude holder of an utterance.

Despite this similarity, these experiments present several points of departure from the literature on true logophoric languages. First, English cannot be said to have true logophors (in Culy's sense), since a single morphological form would be shared between reflexivity and logophoricity. Setting this aside, English is not behaving in the strict fashion described by Culy for West African languages. Languages with true logophoric morphology reportedly allow it in some embedded contexts, and strictly forbid it in others. In contrast, English appears to permit long-distance reference in many embedded contexts, using attitude verbs in a gradient fashion to assist in the assignment of perspective taking. So while English reflexives and West African logophors both seek non-local perspective centers, and both rely on attitude verbs in doing so, the manner in which these verbs are employed is importantly different. For the former, reference in some contexts is simply forbidden, while for the latter it is merely dispreferred, or else more difficult.

This then raises the question of how best to represent the effect of attitude verbs in English. If attitude verbs aren't grammatically assigning discourse roles to their subjects, then in what manner are they influencing the choice of perspective center? Regrettably, the answer to that question likely lies outside the scope of this dissertation, relying on information about how humans track perspective in discourse more generally. However, I can offer a few tentative speculations on this topic. Suppose that humans, generally, track perspective when engaged in conversation. This seems like a reasonable assumption given that it is generally necessary to know the perspective with respect to which one should be interpreting a proposition. Given this general communicative desideratum, it seems entirely plausible that linguistic structures might be co-opted to assist in this endeavor. In fact, in logophoric languages, this seems to be exactly what has happened. Moreover, these languages seem to draw even finer distinctions, which, eventually, are ossified into grammatical fact, resulting in the linguistic encoding of highly specific pieces of perspective. Under this view, English may be at the very early stages of this process. Reflexive pronouns may have been selected as the vehicle for codifying perspective (as they have been in other languages), and speakers may be intuiting that some referents are more plausible as perspective centers than others. In the case of speech verb subjects, they correspond to all three discourse primitives described by Sells. In contrast, perspective verbs subjects at best refer to the role of PIVOT. Given this, the speaker might induce a preference for reference to speech verb subjects, which make more complete perspectives holder, even though this preference is not yet grammatically encoded. In fact, this is one way of understanding the difference between what Culy calls "true" and "mixed"

logophor languages. Perhaps speakers of languages like Mandarin, Japanese, and Icelandic have begun to make distinctions among kinds of perspective holder, but have not yet lexically encoded this information in the verbs used to embed logophoric reference. From this perspective, there is a clear cline of langauge evolution, from langauges like English (with relatively primitive reference to perspective holder), to Mandarin and Japanese (which make further distinctions in the classes of perspective holder), to languages like Ewe, which have fully grammaticized the embedded environments in which logophoric reference can occur. At present this remains a fairly speculative line of reasoning, but it is one to which we will return in Chapter 5.

CHAPTER 3
PERSON BLOCKING IN REFLEXIVE PROCESSING

One very clear consequence of Chapter 2 was the impact of perspective taking on sensitivity to lure-referents in reflexive comprehension. What started as an investigation of the impact of attitude verb-type quickly turned into an exploration of the manner in which the perspectival properties of linguistic units in a sentence impacted the preferred reference of a reflexive pronoun. In light of Experiment 2a, it seems that sensitivity to lures is conditioned on the degree to which the lure can act as the perspective center of the utterance. This makes the prediction that attention to lures should be attenuated in the presence of targets which inherently act as strong centers of perspective.

One such situation is indexical ponouns, which have been argued elsewhere to be strong attractors for the role of perspective center (Huang & Liu, 2001; He & Kaiser, 2012). If this is true, then we would expect indexical pronouns to provide another barrier to non-local reference, much in the same way that perception verbs did in Experiment 1. Chapter 3 is aimed at this addressing this question with two experiments. Experiment 3 tests the prediction that indexical pronouns (e.g. *I*, *you*) act as automatic perspective centers, thereby preventing other, non-indexical referents from anteceding an embedded reflexive. Experiment 4 pushes this further to show that the interference imposed both indexicals is not a product of animacy, but rather of person.

The chapter is organized as follows: Section 3.1 gives an account of the impact of indexical pronouns on putatively logophoric pronouns cross-linguistically. Section 3.1 then presents the results of Experiment 3, demonstrating that indexical pronouns do indeed impede access to the lure referent. Section 3.3 presents the results of Experiment 4, demonstrating that the impedance of indexical pronouns is not a matter of animacy, necessarily, but rather a function of their privileged status as perspective holders. Finally, Section 3.4 brings these results together into a discussion of the role, and representation of perspective in reflexive comprehension.

3.1 Person blocking and logophoricity

In Chapter 2, we saw that logophoric pronouns typically target the perspective center of an utterance. While languages vary in how this goal is grammaticized, the underlying generalization remains largely intact. We also saw evidence that manipulating the presence/absence of competition for the role of perspective center can impact the likelihood of attending to non-local antecedents. When lure and target antecedents compete as perspective holders, some attitude verbs appear to make the lure a more desirable perspective target. This observation makes an interesting prediction: if perspective-holding is critical for producing sensitivity to lure referents, then it should, in principle, be possible to negate this sensitivity by making lures comparatively *less* attractive as perspective centers. That is, it should be possible to manipulate targets such that they are more tempting perspective holders, thereby reducing sensitivity to lure referents.

Cross-linguistically, this appears to be true, perhaps nowhere more clearly than in the case of the Mandarin reflexive *zji*. As mentioned above, *ziji* falls into the class of "mixed" logophors (to borrow Culy's term), apparently leading a double life as reflexive and logophoric pronoun. Notably, *zji* can always be locally bound, as in (89a). However, it may also be bound at a distance, meaning that sentences like (89b) are ambiguous, with both referents capable of anteceding the embedded anaphor. *Ziji* thus poses a problem for the standard binding theory. It can be locally bound, and therefore should be reflexive. However, it can also be bound outside its governing category (roughly: the minimal phrase containing the anaphor and a subject), a property not generally associated with reflexivity.

(89) a. Lisi$_i$ hen ziji$_i$.

Lisi$_i$ hates SELFi
Lisi hates herself

b. Zhangsan$_i$ renwei Lisi$_j$ hen ziji$_{i/j}$

Zhangsan$_i$ thinks Lisi$_j$ hates SELF$_{i/j}$
Zhangsan thinks Lisi hates himself/herself

Recognizing this problem, many early approaches to *ziji* sought to refine the definition of "governing category" for Mandarin, or else, reconfigure binding theory entirely (Xue, Pollard, & Sag, 1994). However, Huang and Liu (2001) argued that neither approach is necessary. Instead, they claimed that long-distance interpretations of *zji* were fairly restricted, and followed from an independent, logophoric use of the anaphor. In support of this position, they note that, when long-distance bound, *ziji* refers to a discourse SOURCE (in the sense of Sells (1987)), and that in

these cases it must be interpreted *de se*. That is, sentences like (90) are only licit if Zhangsan in fact recognizes the stolen purse as his own. This *de se* requirement has also been argued to hold of true logophors[1], as in Bafut (Kusumoto, 1998), Yoruba (Anand, 2006), and Tangale (Haida, 2009). Given these constraints on distribution of long-distance interpretations, Huang and Liu (2001) suggest that *ziji* is, in fact, lexically ambiguous, with one use corresponding to a traditional Principle A anaphor, and the other functioning like a logophoric pronoun. Handily, this obviates the need for amending or reconfiguring the standard Binding Theory approach.

(90) Zhangsan$_i$ shuo pashou tou-le ziji$_i$-de pibao
Zhangsan said pickpocket steal-PERF SELF$_i$-GEN purse
Zhangsan$_i$ said the pickpocket stole his$_i$ purse.

Building on this premise, the authors then use long-distance *ziji*'s logophoric nature to explain the otherwise puzzling set of data in (92) and (93). In (91a), we see as before that embedded *ziji* is ambiguous: capable of referring to either the embedded, or the matrix subject. This remains true when the matrix subject is replaced with a first or second person (indexical) pronoun, as seen in (91b)[2]. However, if the indexical pronoun intervenes between embedded *ziji* and its intended long-distance referent, the ambiguity vanishes and *ziji* must be bound locally, as seen in (92). Interestingly, this is true even if the indexical is not, itself, capable of anteceding *zji*, as seen in (92b)[3]. Thus, it seems that the mere presence of an intervening indexical pronoun is sufficient to prevent *ziji* from taking a long-distance antecedent. Similarly, a singular, third person local referent prevents embedded *ziji* from referring to a non-local plural antecedent, though reference to a non-local, singular referent over a plural local antecedent is permitted. This contrast is given in (93). Collectively, these phenomena are referred to as "blocking", as an intervening referent appears to "block" a non-local interpretation of *ziji*. For the present, we will focus more narrowly on "person blocking", as seen in (92), though number blocking will resurface in later discussion.

(91) a. Zhangsan$_i$ danxin Lisi$_i$ hui piping ziji$_{i/j}$.
Zhangsan$_i$ worry Lisi$_i$ will criticize SELF$_{i/j}$
Zhangsan worries that Lisi will cricize him/herself.

---

[1] cf. Recent work by Pearson (2015), who provides evidence that *yè* in Ewe is not always interpreted *de se*.

[2] Note that *zji* is person and number ambiguous, and so can grammatically refer to either indexical, or third-person antecedents.

[3] *Ziji* is obligatorily subject-oriented, and so cannot refer to the first person pronoun in this sentences because it is an object.

b. Wo/ni danxin Lisi hui piping ziji.

I/you$_i$ worry Lisi$_j$ will criticize SELF$_{i/j}$
I/you worry that Lisi might criticize me/you/herself.

(92) a. Zhangsan$_i$ danxin wo/ni$_j$ hui piping ziji$_{*i/j}$.

Zhangsan$_i$ worry I/you$_j$ will criticize SELF$_{*i/j}$
Zhangsan worries that I/you might criticize my/yourself.

b. Zhangsan$_i$ gaosu wo$_j$ Lisi$_k$ hen ziji$_{*i/*j/k}$.

Zhangsan$_i$ tell me$_j$ Lisi$_k$ hate SELF$_{*i/*j/k}$
Zhangsan told me that Lisi hates herself.

(93) a. Lisi$_i$ zhidao tamen$_j$ chang piping ziji$_{i/j}$.

Lisi know they often criticize SELF
Lisi knows that they often criticize her/themeselves.

b. Tamen$_i$ zhidao Lisi$_j$ chang piping ziji$_{*i/j}$.

They$_i$ know Lisi$_j$ often criticize SELF$_{*i/j}$
They know Lisi often criticizes herself.

Huang and Liu (2001) contend that this pattern of judgments can be explained by appealing to *ziji*'s logophoric nature. Specifically, they suggest that logophoric *ziji* enters the discourse representation as a pronoun anchored to a sentence internal SOURCE. Thus, when *ziji* is assigned a long-distance interpretation, the non-local referent must have been assigned the SOURCE role. Indexical pronouns, however, obligatorily anchor the SOURCE role to a sentence *external* referent: either the speaker, or the addressee. Consequently, in sentences containing an indexical, other, third-person referents cannot be assigned the SOURCE role without creating a conflict of perspective (under the assumption that there can be only one SOURCE for an utterance). Working with these assumptions, the pattern in (92) can be readily explained. In order for *Zhangsan* to non-locally antecede *ziji* in (92a), he would have to be the internal SOURCE of the sentence. However this sentence contains an indexical pronoun, which obligatorily binds the SOURCE role. Thus, *Zhangsan* cannot fill the role necessary to act as an antecedent for embedded *ziji*, and person blocking arises.

In addition, this system handily explains the fact that the indexical pronoun need not, itself, be able to bind *ziji* to induce blocking, as seen in (92b). The problem with indexicals is that they bind the SOURCE role, not that they out-compete other referents as direct binders for *ziji*. Finally, we also have a good understanding of the fact that non-intervening indexical pronouns, as in (91) do not unambiguously bind *zji*. In these cases, the indexical can act as a non-local binder because it is, inherently, a SOURCE. However, *ziji* is also a standard anaphor, and therefore capable of local

binding irrespective of the SOURCE of the utterance. Therefore, in cases like (91), the indexical can bind logophoric *ziji* at a distance, while the local subject is capable of binding anaphoric *ziji* in the usual manner.

This pattern of behavior, and the associated explanation[4], provide precisely the tool needed to test the prediction laid out at the beginning of this section. If indexical pronouns act as tempting perspective centers (perhaps because they obligatorily bind SOURCE), then we should expect them to decrease the availability of lures as logophoric antecedents. Specifically, this leads us to expect lure referents in sentences like (94) to be less accessible when the embedded subject is a first or second person indexical.

(94) The actress said that $\left\{\begin{matrix}\text{it}\\ \text{I}\end{matrix}\right\}$ horribly misrepresented herself in the article.

On the logophlexives hypothesis, *actress* in (94) should be less available as an antecedent for the reflexive when there is a local first person pronoun *I* than when there is a local 3rd person pronoun *it*. Here, *I* will act as an obligatory SOURCE, preventing the lure from acting as a sentence internal perspective holder. In contrast, if lure referents are available when the local subject is a particularly poor morphosyntactic match for the reflexive (e.g. Parker & Phillips, 2017), then we expect lure referents to be equally available when the local subject is *it* or *I*. Experiment 3 tests these predictions with two acceptability judgment surveys, and one eye-tracking while reading study.

3.2 Experiment 3: Person blocking in reflexive comprehension

Experiment 3 was designed to test for person-blocking effects in English reflexive processing. To the extent that we observe such effects, we have further evidence implicating a logophoric source for Principle A fallibility. To assess this, pairs of context/test of sentences were manipulated as shown in (95). Test sentences always contained a reflexive in the embedded direct object position of a bi-clausal structure, and the embedding verb was always a speech predicate to optimize the lure's status as a potential logophoric antecedent. As in Experiments 1 and 2, the gender of the matrix subject in the test sentence was then manipulated so that it either matched,

[4]The wary reader might, at this point, be reasonably concerned with the characterization of person blocking as a fact about logophoricity, broadly construed. Indeed, while Huang and Liu (2001) provide a compelling analysis of the phenomenon in terms of logophoric principles (i.e. Sells' discourse roles), theirs is hardly the only analysis on the market. Indeed, more recent approaches have favored treating long-distance *ziji* as a shifted indexical, rather than a logophoric pronoun (Anand, 2006). Moreover, while blocking seems robust in Mandarin, it is not clear that it is a general property of logophoricity, and is not, to my knowledge, discussed in the literature on true, West African logophors. These concerns, and their implications for the data to follow, will be addressed more thoroughly at the conclusion of this chapter.

or mismatched the embedded reflexive (LURE: ±*match*). Finally, the embedded subject of the test sentence was manipulated so that it was either a proper name, a third person inanimate pronoun, or a first person pronoun (TARGET: *Name*, *it*, *I*). Concomitant with this manipulation, the context sentence was manipulated so that it introduced an appropriate referent for the embedded subject in the test sentence. Contexts always introduced a generic plural noun phrase (paired with *Name* test sentences), an inanimate noun phrase (paired with *it* test sentences), or a first person pronoun (paired with *I* sentences). These manipulations resulted in two factors (LURE: ±*match*; TARGET: *Name, it, I*), which, when fully crossed, produced six conditions. As in previous experiments, separate acceptability judgment and eye-tracking while reading studies were carried out on the same set of items.

(95) **Context:** { Some movie critics / The salacious tabloid / I } said some unflattering things about Hollywood icons.

Target: The { actress / actor } | said that| { Joanna / it / I } horribly misrepresen|ted herself| in the article|...

In this design, person-blocking would be realized as a substantial lure match effect on reading times and judgments of the embedded reflexive after *it*, but not after *I*. In addition, it is possible, but unlikely, that lure-match effects would be observed for *Name* conditions, which were always a target-match for the reflexive. Given that we failed to find an effect of lure match for normatively grammatical sentences in our previous experiments (as did (Parker, 2014)), it seems likely that they will not manifest here, either.

3.2.1 Acceptability judgments

3.2.1.1 Experiment 3a: First person blocking

53 self-reported Native English speakers were recruited via Amazon Mechanical Turk to rate sentences on a 1-7 scale (1=*completely unnatural*; 7=*completely natural*). Sentence pairs were presented above the scale, with each sentence labeled *context* and *target* respectively. Participants were instructed to read both sentences carefully, but to base their judgment on the naturalness of the test sentence only. Presentation of filler and experimental items was fully randomized. As with Experiments 1a, and 2a the experiment was coded and hosted on the Ibex Farm server.

3.2.1.2 Materials

36 context/test pairs patterned on (95) were created in a latin-square design. 72 additional context/test pairs were created as filler sentences, and randomly interleaved with the experi-

mental items. Fillers were composed of a mixture of sentences embedding indirect questions (24 sentences) and grammatical instances of first-person and third-person inanimate pronouns as embedded subjects (48 sentences). This latter measure was particularly important in preventing participants from adopting a strategy whereby the person/animacy features of the embedded subject could act as a reliable cue to sentence acceptability. In total, 22% of the materials were normatively ungrammatical.

3.2.1.3 Analysis

A mixed effects model was fit to sentence ratings for the experimental items. The factors LURE, TARGET and their interaction were used as fixed effects, with uncorrelated random slopes and intercepts assigned to subjects and items. The two-level factor LURE was sum-coded ($+match{=}1$, $-match{=}-1$) to test for an effect of lure match. The three-level factor TARGET was broken down into two separate contrasts, one probing for an effect of target match (TARGET: *Name*$=1$, *it*$=-.5$, *I*$=-.5$), and the other for an effect of embedded person (PERSON: *Name*$=0$, *it*$=1$, *I*$=$-1). Eight participants who reported exposure to an East Asian or West African language, participation in a previous study about reflexives, or being older than 55 were excluded prior to analysis (n=45).

3.2.1.4 Results

A summary of results is given in Table 3.1, with the mixed effects model coefficients reported in Table A.6 in the appendix. Unsurprisingly, participants rated sentences with a matched target antecedent (*Name* conditions) significantly better than those with a mismatched target antecedent. This was realized as a significant main effect of TARGET ($\hat{\beta}$=1.29, t=8.35). In addition, there was a main effect of PERSON, indicating that sentences with *I* as an embedded subject were rated less natural than those with *it* ($\hat{\beta}$=0.23, t=3.01). The effect of LURE match also reached significance ($\hat{\beta}$=0.16, t=4.34), indicating higher ratings for sentences with reflexives which matched the lure referent. These main effects were qualified by a TARGET$\times$LURE interaction: sentences with mismatched target antecedents were given higher acceptability ratings when the reflexive matched the lure ($\hat{\beta}$=-0.21, t=4.61). Nested pairwise comparisons revealed a significant effect of lure match on both *it* ($\hat{\beta}$=-0.65, t4.98) and *I* ($\hat{\beta}$=-0.39, t=2.96) sentences, but not *Name* sentences, confirming the interpretation of the TARGET$\times$LURE interaction.

These results demonstrate a substantial effect of the lure on the perceived acceptability of the embedded reflexive. When the reflexive mismatched the target antecedent (*it* and *I* conditions), acceptability was significantly increased by a feature-matched lure. This effect was numerically

Table 3.1. Experiment 3a: Mean by-subject naturalness ratings (standard error in parentheses)

	Lure	
Target	*+match*	*−match*
Name	5.13 (0.17)	5.24 (0.16)
It	3.80 (0.20)	3.13 (0.20)
I	3.20 (0.20)	2.82 (0.19)

larger for *it* sentences than for *I* sentences, suggesting a trend towards a person-blocking effect, but the interaction of LURE and PERSON did not reach significance. Experiment 3a thus provides incomplete evidence of person blocking in English reflexive reference. We turn now to Experiment 3b, a replication and extension of the current study which tested for effects of person blocking using second person pronouns.

3.2.1.5 Experiment 3b: Second Person Blocking

The materials for Experiment 3b were identical to those in Experiment 3a, except that all instances of first-person pronouns were replaced with second person pronouns. 54 self-reported Native English speakers were recruited via Amazon Mechanical Turk. The experiment procedure and analysis was identical to that of Experiment 3a (substituting *you* in place of *I* in the analysis). Four participants was excluded from analysis using the same rejection criteria as before (n=50).

3.2.1.6 Results

A summary of results is given in Tables 3.2, with the mixed effects model analysis given in A.7 in the appendix. As in Experiment 3a, sentences in which the target antecedent and reflexive matched (*Name* conditions) received higher ratings than the target mismatch sentences, reflected in a significant main effect of TARGET ($\hat{\beta}$=1.40, t=9.42). There were also main effects of LURE ($\hat{\beta}$=0.20, t=5.42), reflecting higher ratings when the matrix subject matched the reflexive, and PERSON ($\hat{\beta}$=0.19, t=4.2), indicating lower ratings for *you* sentences than *it* sentences. As in Experiment 3a, these main effects were qualified by a significant TARGET×LURE interaction: lure match impacted ratings for target-mismatch sentences more than target-match sentences ($\hat{\beta}$=-0.23, t=-4.16). However, the model also revealed a significant PERSON×LURE interaction, indicating a larger effect of lure match on *it* sentences than on *you* sentences ($\hat{\beta}$=0.12, t=2.55). This interpretation was borne out by nested pairwise comparisons, which revealed a significant lure match effect for *it* sentences ($\hat{\beta}$=-0.89, t=7.37) which was larger than that for *you* sentences ($\hat{\beta}$=-0.38, t=3.16).

These results directly reflect person blocking in English reflexive reference. While reflexives in both *it* and *you* sentences showed improved acceptability in the presence of a feature matched

Table 3.2. Experiment 3b: Mean by-subject naturalness ratings (standard error in parentheses)

	Lure	
Target	*+match*	*−match*
Name	5.47 (0.14)	5.55 (0.12)
it	4.05 (0.22)	3.15 (0.21)
you	3.41 (0.21)	3.00 (0.20)

matrix subject, the presence of the second person pronoun substantially decreased the degree of amelioration. With these striking results surfacing in off-line measures, we turn now to the eye-tracking while reading study to assess the impact of indexical pronouns on real-time reflexive interpretation.

3.2.2 Eye-tracking while reading

40 monolingual, English-speaking UMass undergraduates participated for extra credit in introductory linguistics and psychology courses. Details of participant exclusion are given in the analysis section.

3.2.2.1 Materials

The same items from Experiment 3a (including fillers) were used in the eye-tracking study. Every item was followed by a two-alternative choice comprehension question. Half the comprehension questions probed the contents of the context sentence, while the remainder probed aspects of the test sentence other than the reference of the reflexive (by-subject question accuracy mean=87%).

3.2.2.2 Procedure

The same EYELINK system and setup as in Experiment 2b were used in Experiment 3c.

3.2.2.3 Analysis

As in the previous eye-tracking studies, first pass, go-past, and total reading times at the reflexive and spillover regions were analyzed. The reflexive region was again simply the embedded reflexive, extended three characters to the right to accommodate high skipping rates (21% vs. 10% before and after extension, respectively). The spillover region contained the material to the right of the reflexive, up through the first content word.

Trial and artifact rejection criteria were identical to those used in Experiments 1b and 2b. four participants were excluded from analysis due to excessive data loss (more than 25%). In

addition, inordinately long first pass, and total time values were removed from the data (>2000ms or 4000ms, respectively).

The same mixed effects model structure used in the acceptability judgment task was adopted for each of the three fixation duration measures. This included the main effects and interactions of LURE(+*match*=-1, −*match*=1), and the helmert coded factor TARGET (TARGET:*Name*=-1, *it/I*=.5; PERSON: *Name*=0, *it*=-1, *I*=1). Random slopes and intercepts were assigned to each fixed effect by subject and item, excluding correlations among the random effects.

3.2.2.4 Results

A summary of results at the reflexive, and spillover regions is presented in Table 3.3. Effects in go-past and total reading time at the both regions are depicted in Figure 3.1. The results of the mixed effect models fit to these data are given in Table A.8 in the appendix.

Table 3.3. Experiment 3c: By subject means for first pass, go-past, and total times at the reflexive and spillover regions (standard errors in parentheses)

		Reflexive			Spillover		
Target	**Lure**	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
Name	*+match*	310 (17)	360 (21)	426 (28)	382 (18)	473 (33)	514 (30)
	-match	291 (11)	368 (31)	407 (20)	377 (17)	503 (51)	535 (30)
It	*+match*	280 (11)	360 (26)	427 (22)	365 (18)	680 (59)	564 (33)
	-match	322 (16)	471 (38)	543 (37)	382 (21)	720 (91)	628 (57)
I	*+match*	309 (15)	517 (49)	541 (33)	402 (22)	664 (58)	662 (40)
	-match	351 (18)	435 (26)	634 (43)	417 (27)	918 (89)	759 (59)

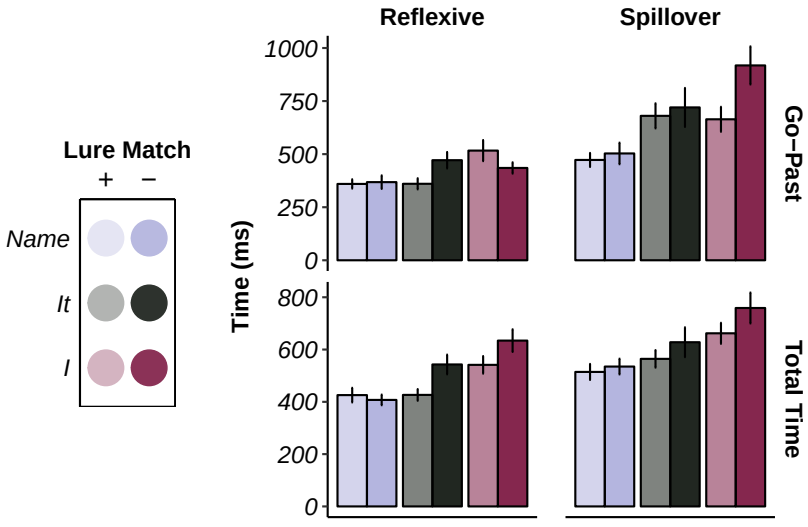


Figure 3.1. Experiment 3c: Mean by-subject go-past and total time reading measures at the embedded reflexive region. Error bars represent standard error

In first pass and total reading time at the reflexive region, we observe main effects of PERSON (first pass:$\hat{\beta}$=16, t=3.19; total time: $\hat{\beta}$=50, t=3.88), and LURE (first pass: $\hat{\beta}$=10, t=2.45; total time: $\hat{\beta}$=33, t=3.48), indicating longer reading times for reflexives following first-person pronouns, and faster reading times for reflexives with feature-matched lures. In go-past and total reading time, there was also a a main effect of TARGET (go-past: $\hat{\beta}$=52, t=3.32; total time: $\hat{\beta}$=82, t=6.17), corresponding to longer reading times for reflexives which mismatched the target antecedent. These main effects were qualified by a significant TARGET×LURE interaction in first-pass and total reading time (respectively: $\hat{\beta}$=19, t=3.10; $\hat{\beta}$=38, t=2.43). The positive coefficient for the interaction indicates that feature matched lures facilitated reading of target-mismatched reflexives (*it* or *I*), but not target-matched reflexives (*Name*). In contrast, there was an interaction of PERSON×LURE in go-past times at the reflexive ($\hat{\beta}$=-47, t=3.66). The negative coefficient for the PERSON×LURE interaction indicates greater lure-match facilitation for reading target-mismatched reflexives when the embedded subject was *it*, relative to *I*.

Nested pairwise comparisons testing the lure match effect confirm these interpretations. In all three measures there was a significant lure-match facilitation effect for reflexives in "it" sentences (first pass: $\hat{\beta}$=23, t=3.09; go-past: $\hat{\beta}$=57, t=3.06; total-time: $\hat{\beta}$=59, t=3.78). For sentences in which the embedded subject was "I", there was a significant lure-match facilitation effect in first pass ($\hat{\beta}$=17, t=2.37) and total reading time ($\hat{\beta}$=45, t=2.96). However, in go-past reading time there was a significant lure-match *disadvantage* for reflexives following indexicals ($\hat{\beta}$=-37 t=2.02). This flip in the sign of the lure-match effect in go-past reading time is likely partially responsible for the significant PERSON×LURE interaction observed in that measure.

At the spillover region, there was a significant effect of PERSON in both first pass ($\hat{\beta}$=16, t=2.33) and total reading time ($\hat{\beta}$=57 , t=3.70), indicating as before longer reading times for sentences containing an indexical pronoun. In addition, there was again a main effect of TARGET in go-past ($\hat{\beta}$=172, t=5.87) and total times ($\hat{\beta}$=87, t=5.08), representing longer reading times following target-mismatched reflexives. Finally, there was a main effect of LURE in total reading times ($\hat{\beta}$=32, t=2.11), corresponding to faster reading times following lure-matched reflexives. No interactions approached significance at this region. Nested pairwise comparisons revealed a significant lure-match facilitation effect for "I" sentences in both go-past ($\hat{\beta}$=126, t=3.44) and total reading time ($\hat{\beta}$=49, t=2.81). There was a trending lure-match facilitation effect for "it" sentences in total reading time, but it failed to reach significance ($\hat{\beta}$=33, t=1.89). No other effects approached significance.

3.2.2.5 Summary of fixation duration analyses

Analyses of fixation duration-based measures in Experiment 3c revealed somewhat mixed evidence of person blocking in reflexive comprehension. First, in go-past reading times at the reflexive, we saw reasonable evidence of person blocking in action: reflexive reading was facilitated by feature matched lures, but not when the local subject was an indexical pronoun. Unfortunately, this effect did not generalize to other measures or regions. In first pass and total reading times at the reflexive, and in go-past and total-times at the spillover, we observed substantial lure match facilitation associated with both *it* and *I* sentences, indicating no particular blocking effect for indexicals. Before abandoning person blocking as an outcome in Experiment 3c, however, let's assess the cumulative progression data.

3.2.2.6 Cumulative progression analysis

As in previous experiments, a cumulative progression analysis of fixations past the reflexive was conducted. Once again, the cumulative progression curves were statistically evaluated using a cluster mass permutation test, assuming an α cutoff of 0.9 for cluster identification. This test evaluated difference-of-differences equivalent to the TARGET×LURE and PERSON×LURE interactions in the mixed effects model. These interaction terms were the result of the nested pairwise comparisons shown in Figure 3.2, comparing the relative effect of lure match (*match*−*mismatch*) on target match/mismatch sentences (*Name* vs. *it*/*I*), and on first-person/third-person intervention sentences (*it* vs. *I*).

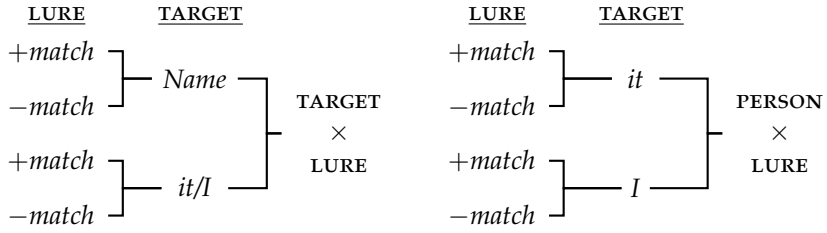


Figure 3.2. Experiment 3c: Nested pairwise comparisons calculated for separate cluster mass permutation tests of TARGET×LURE and PERSON×LURE. Each parent node is the difference (top-bottom) of its daughter nodes

The cumulative progression analysis exhibited a qualitatively similar pattern to the fixation duration measures. Cluster mass permutation tests revealed an early PERSON×LURE interaction which was significant by subjects (490–810ms, $p<.05$), and by items (490–660ms, $p<.05$). In addition, there was an early interaction of TARGET×LURE which was significant by subjects (550–620,

$p<.05$), and a late-going effect which was significant by items (2150–4000ms, $p<.05$). This is analogous to the difference between the effects observed in go-past and total reading times: the lure-match advantage was reduced in an early time window for *I* sentences relative to *it* sentences (similar to go-past times), but there was a generic lure-match advantage in a later time window, regardless of person of the embedded subject (similar to total reading times).

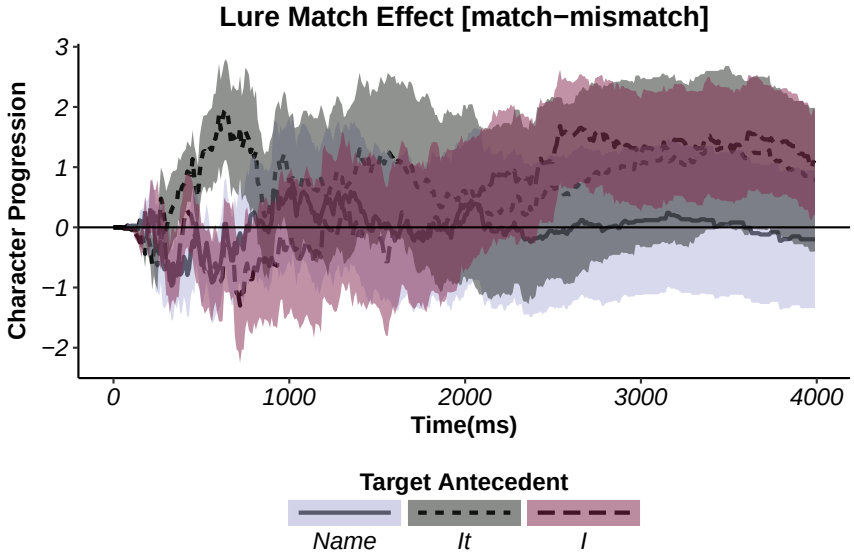


Figure 3.3. Experiment 3c: Mean by-subject lure-match effects (match-mismatch) in characters from first fixation on the reflexive. Error bars represent a bootstrapped 95% confidence interval

3.2.2.7 Summary

Overall, Experiment 3c provided consistent evidence of person blocking, albeit a more gradient, transitory version of the effect. In two dependent measures (go-past reading times at, and cumulative progression past, the reflexive region) we saw evidence of early lure-match facilitation effects for *it* sentences which was absent for *I* sentences. However, after this initial person-blocking response, there was more or less homogeneous lure-match facilitation for both *it* and *I* sentences (as seen in fixation duration measures at the spillover region and at a late cumulative progression time window). If this characterization is correct, then we have evidence that person blocking may be active in the initial stages of antecedent resolution, but in later, plausibly repair-driven, stages.

Unfortunately, there is one datum inconsistent with this interpretation of these results: there was a significant lure-match facilitation found for reflexives following *I* in first pass reading at the reflexive region. Given this, it seems disingenuous to claim that indexicals act as an early deterrent of lure-match facilitation. That said, it is worth noting that even when lure-match facilitation

was found in indexical conditions, it still did not completely ameliorate the target-match violation. That is, for target-mismatch sentences containing *it*, lure-match facilitation rendered reading times more or less on-par with those of the grammatical, target-match controls. In contrast, indexical sentences were associated with an overall penalty (realized as a main effect of PERSON), which was never completely obviated. This suggests that while indexicals don't prevent comprehenders from entertaining lure referents, they do make that prospect unappealing. Put in terms of the discussion of *ziji*, this may be an indication that comprehenders are capable of entertaining conflicting SOURCES, but they dislike doing so. If so, we should expect similar results to obtain in Mandarin, even though person blocking is a recognized, grammatical fact of that language.

3.2.3 Discussion

On the whole, Experiments 3 presents consistent, if somewhat messy evidence of person blocking in English reflexive comprehension. Notably, we observed three strong instances of blocking across two experiments. In experiment 3b, we observed a substantially reduced lure-match facilitation effect for target-mismatched reflexives following the indexical *you*. In Experiment 3c, we saw in two dependent measures (go-past reading, and cumulative progression) that attention to lures was attenuated in the presence of an indexical pronoun. Also suggestive was the numerical trend observed in Experiment 3a, where the lure-match effect was, at least numerically, stronger for *it* sentences than for *I* sentences. Finally, we saw that even when lure-match facilitation was observed in indexical conditions, it was accompanied by an overall penalty associated with the indexical. In sum, while these effects do not reflect the tidy, categorical phenomena reported in Mandarin[5], they do represent evidence that indexicals impede reference to non-local referents, even if they don't categorically disallow it. These effects are expected under the logophlexives hypothesis if: (*i*) attention to lures is modulated by their viability as logophoric antecedents, and (*ii*) indexical pronouns are preferentially interpreted as the perspective from which an utterance is interpreted.

Interestingly, all three studies studies consistently indicated that person blocking is a gradient, rather than categorical phenomenon. In all three experiments, the presence of a first/second person pronoun reduced, but did not entirely eliminate, the impact of lure referents. We cannot,

[5]As an aside, it isn't at all clear that the effects reported in Mandarin are nearly as categorical as they are sometimes reported in the literature. At least some theoretical sources report substantial disagreement in judgments over some critical examples (Anand, 2006), while recent experimental work has shown the effect to be quite gradient and, if anything, stronger with second-person pronouns (He & Kaiser, 2012). Given this, it is not at all clear that the results of Experiment 3 are outside the realm of expected behavior for person blocking.

at present, offer a definitive explanation of this finding, though one plausible interpretation lies in the manner in which indexical pronouns shift perspective. Recall that person blocking effects are explained by appealing to two facts: (*i*) long-distance reflexives are logophoric and must refer to a perspective center, and (*ii*) that indexical pronouns act as perspective centers. This account predicts categorical person blocking only in so much as indexical pronouns obligatorily act as perspective centers. If, on the other hand, indexicals only probabilistically anchor the discourse perspective, then we would expect gradient person blocking effects. We might then explain the stronger blocking effects observed for second person by positing that *you* is more likely to act as a perspective center (for some as yet unclear reason). This explanation is similar to the explanation offered by He and Kaiser (2012) for the variable person blocking effects they observe in Mandarin.

In this sense, Experiment 3 joins Experiment 2 in demonstrating that some previously categorical phenomenon (in that case, the effect of verb type) was rather more gradient for English than the logophlexives hypothesis would (necessarily) lead us to believe. As a result, a picture emerges in which attention to lures is modulated by the degree to which speakers probabilistically assign them the role of perspective center. Attitudinal contexts and indexicality influence this decision, but do not seem to categorically constrain it. This theme will occupy some of the discussion section at the end of this chapter, and animate much of the analysis presented in Chapter 5.

3.3 Experiment 4: De-confounding person and animacy

One possible confound in Experiment 3 is that the target-mismatch, un-blocked condition was inanimate, while the blocking configuration necessarily involves an animate target antecedent (*I*/*you*). Given the results of Experiment 2, this is potentially highly concerning: we know already that inanimate targets make the lure much more available than it otherwise would be. Thus, it is possible that the person-blocking effects observed in Experiment 3 (such as they were) were really a function of animacy, rather than person. Put differently, Experiment 3 could have produced the *intended* effect independent of the blocking potential of indexical pronouns. Experiment 4 controls for this confound by replacing the embedded subject in target-mismatch, un-blocked conditions with the plural third-person pronoun *they*. A sample paradigm is given in (96). If the effects observed in Experiment 3 were in fact due to the intervention of an indexical, and not the inanimacy of the target in the non-blocking conditions, then we should observe the same qualitative pattern in this experiment. As in previous experiments, separate acceptability judgment and eye-tracking while reading experiments were conducted.

(96) **Context:** $\left\{\begin{array}{c}\text{Some movie critics}\\ \text{I}\end{array}\right\}$ said some unflattering things about Hollywood icons.

Test: The $\left\{\begin{array}{c}\text{actress}\\ \text{actor}\end{array}\right\}$| said that| $\left\{\begin{array}{c}\text{Joanna}\\ \text{they}\\ \text{I}\end{array}\right\}$ horribly misrepresen|ted herself| in the article|...

3.3.1 Acceptability judgments

76 self-reporting native English speakers were recruited via Amazon Mechanical Turk and compensated $5 for their participation. Prior to analysis, 18 participants were excluded for reporting exposure to an East-Asian or West African language, due to age (older than 55), or for indicating prior participation in a study about reflexive pronouns. The remaining 58 participants were between the ages of 21 and 53 (median age: 32).

3.3.1.1 Materials

The materials of Experiment 4 were adapted from those of Experiment 3. In place of the "it" conditions, this experiment used the generic, plural referent of the context sentence as a referent for the third person plural pronoun *they*. Filler sentences were adapted so that what had been "it" fillers in Experiment 3 served as "they" fillers in Experiment 4. All other aspects of the materials in Experiment 4 were identical to those used in Experiment 3.

3.3.1.2 Analysis

A linear mixed effects regression was fit to sentence ratings for the experimental items. In all major respects, this model was identical to the one fit to the data in Experiments 3a-b, with the exception that a *they* vs. *I* contrast replaced the *it* vs, *I* contrast. The two-level factor LURE was sum-coded ($+$*match*=1, $-$*match*=-1), while the three-level factor TARAGET was helmert coded to tests for independent effects of target match (TARGET: *Name*=1, *they/I*=-.5) and embedded person (PERSON: *Name*=0, *they*=1, *I*=-1). As before, correlations among the random effects were not included in the model.

3.3.1.3 Results

A summary of by-subject mean naturalness ratings is given in Table 3.4. A table of fixed effects taken from the model fit to these data is presented in Table A.9 in the appendix. As in previous studies, this model revealed significant main effects of TARGET ($\hat{\beta}$=1.36, t=10.28), LURE ($\hat{\beta}$=0.10, t=3.36), and PERSON ($\hat{\beta}$=0.12, t=2.78), indicating that sentences were rated better when the reflexive matched the target, when it matched the lure, and when the target wasn't an indexical.

These main effects were qualified by a significant TARGET×LURE interaction, indicating that lure-match had a greater effect on target-mismatch sentences than on target-match sentences ($\hat{\beta}$=-0.13, t=2.82). The PERSON×LURE interaction did not approach significance. Pairwise comparisons of the effect of lure nested within levels of TARGET found a significant lure-match advantage for reflexives following "they" ($\hat{\beta}$=-0.40, t=3.63), as well as those following "I" ($\hat{\beta}$=-0.26, t=2.39).

Table 3.4. Experiment 4a: Mean by-subject naturalness ratings (standard error in parentheses)

	Lure	
Target	*+match*	*−match*
Name	5.27 (0.12)	5.31 (0.11)
They	3.57 (0.15)	3.16 (0.16)
I	3.25 (0.17)	2.99 (0.17)

3.3.1.4 Summary

Much like Experiment 3a, these findings present suggestive, but not conclusive, evidence of person blocking in English reflexive reference. In this experiment, we observed substantial lure-match facilitation effects for both *they* and *I* sentences, indicating that feature-matched lures improved the acceptability of target-mismatched reflexives, regardless of person. However, as in Experiment 3a, this effect was numerically larger for *they* sentences than for *I* sentences. Moreover, we again see a main effect of person. Given this, it seems that Experiment 4a joins Experiments 3a-b in suggesting a gradient version of person blocking: indexicals do not categorically block access to lure referents, but they do make such access dispreferred. Keeping this in mind, we turn to Experiment 4b to once more examine person-blocking in on-line reflexive comprehension.

3.3.2 Eye-tracking while reading

44 native, monolingual English speaking UMass undergraduates participated for (extra) credit in introductory linguistics and psychology courses. Details of participant exclusion are given in the analysis section.

3.3.2.1 Materials

The same materials (including fillers) from Experiment 4a were used in Experiment 4b. Every item was followed by a two-alternative choice comprehension question. Half of these questions probed facts about the context sentence, the remainder probed aspects of the test test sentence other than the reference of the reflexive (by-subject question accuracy mean = 85%).

3.3.2.2 Procedure

The same EYELINK system and setup as in prior experiments were used in this experiment.

3.3.2.3 Analysis

As in all previous eye-tracking studies, first pass, go-past, and total reading times at the reflexive and spillover region were analyzed. The reflexive region consisted of the embedded reflexive and the three preceding characters to accommodate high skipping rates (19% vs 10%, before and after extension, respectively). The spillover region contained the material to the right of the reflexive, up through the first content word.

Trial and artifact rejection criteria were identical to those used in Experiments 1b, 2b, and 3c. Four participants were excluded from analysis due to excessive data loss (more than 25%). In addition, inordinately long first pass, and total time values were removed from the data (>2000ms or 4000ms, respectively).

The same mixed effects model structure used in Experiment 4a was adopted for each of the three fixation duration measures. This included the main effects and interactions of LURE (+*match*=-1, −*match*=1), and the helmert coded factor TARGET (*Name*=-1, *they/I*=.5; PERSON: *Name*=0, *they*=-1, *I*=1). Random slopes and intercepts were assigned to each fixed effect by subject and item, excluding correlations among the random effects.

3.3.2.4 Results

A summary of results at the reflexive and spillover regions is presented in Table 3.5. A graphical representation of the effects in go-past and total reading times at both regions is presented in Figure 3.4. A full complete table showing the fixed effect results of the statistical models fit to these data is given in Table A.10 in the appendix. Otherwise, the coefficient and *t* values for significant effects are presented in the body of the text below.

Table 3.5. Experiment 4b: By subject means for first pass, go-past, and total times at the reflexive and spillover regions (standard errors in parentheses)

		Reflexive			**Spillover**		
Target	**Lure**	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
Name	+*match*	283 (11)	357 (21)	424 (28)	390 (21)	555 (62)	565 (44)
	−*match*	298 (11)	339 (14)	394 (20)	370 (18)	448 (26)	514 (26)
They	+*match*	279 (10)	409 (25)	434 (20)	383 (19)	504 (31)	528 (25)
	−*match*	294 (13)	389 (28)	503 (33)	389 (24)	742 (60)	629 (40)
I	+*match*	291 (11)	400 (30)	567 (41)	400 (21)	746 (45)	689 (41)
	−*match*	308 (15)	433 (41)	581 (41)	363 (19)	787 (68)	666 (40)

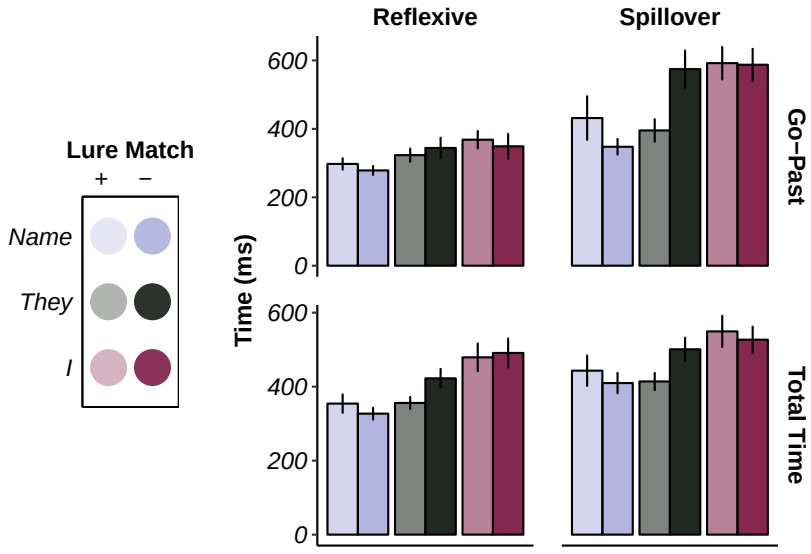


Figure 3.4. Experiment 4b: Mean by-subject go-past and total time reading measures at the embedded reflexive and spillover regions. Error bars represent standard error

At the reflexive region, there was a significant effect of TARGET in go-past ($\hat{\beta}$=41, t=3.22) and total time ($\hat{\beta}$=79, t=5.69) indicating that reflexives which mismatched the local antecedent were read more slowly. In addition, there was a main effect of LURE in first pass reading times ($\hat{\beta}$=8, t=2.02), indicating faster reading times for lure–matched reflexives. Finally, there was a main effect of PERSON on total reading times ($\hat{\beta}$=54, t=2.93), indicating that reflexives following an indexical pronoun were read more slowly overall. No other effects approached significance at this region (all t <1.65). Nested pairwise comparisons testing for a lure-match effect within levels of TARGET revealed a significant lure-match advantage for reflexives following "they" in total reading times ($\hat{\beta}$=39, t=2.43). No other pairwise comparisons approached significance.

At the spillover region, there was again a significant effect of TARGET in go-past ($\hat{\beta}$=139, t=6.25) and total time ($\hat{\beta}$=66, t=4.55). As ever, there were longer reading times when the reflexive mismatched its target antecedent. In both of these measures, there was also a main effect of PERSON, indicating slower reading times following "I" relative to "they" (go-past: $\hat{\beta}$=66, t=2.72; total time: $\hat{\beta}$=50, t=3.58). In addition, there was a main effect of LURE in go-past reading times ($\hat{\beta}$=33, t=2.18), indicating faster reading times following lure-matched reflexives. Also in go-past reading times, both of the TARGET×LURE and PERSON×LURE reached significance (TARGET×LURE: $\hat{\beta}$=69, t=3.22; PERSON×LURE: $\hat{\beta}$=-57, t=2.65). These effects approached significance in total time, but failed to meet it (TARGET×LURE: $\hat{\beta}$=27, t=1.99; PERSON×LURE: $\hat{\beta}$=-31, t=1.88). No other effects approached

significance (all t <1.30). Nested pairwise comparisons revealed a significant lure-match advantage in "they" sentences in both go-past ($\hat{\beta}$=125, t=4.69) and total reading time ($\hat{\beta}$=50, t=2.98). No other pairwise comparisons approached significance.

3.3.2.5 Summary of fixation duration analyses

Descriptively, the fixation duration analyses of Experiment 4b provide stronger support for the existence of person blocking in English reflexive comprehension than has been observed so far. While the critical PERSON×LURE interaction only reaches significance in go-past reading times at the spillover region, this effect was trending towards significance in total time reading time at the spillover (t=1.88), and numerically present in total reading time at the embedded reflexive region. Moreover, pairwise comparisons found substantial lure match facilitation in *they* sentences in all three of these measures, and no evidence of lure-match facilitation associated with *I* sentences.

3.3.2.6 Cumulative progression analysis

As in previous experiments, a cumulative progression analysis of fixations past the reflexive was conducted. The cluster mass test employed was largely identical to the one used in Experiment 3c. This test evaluated difference-of-differences equivalent to the TARGET×LURE and PERSON×LURE interactions in the mixed effects model. These interaction terms were the result of the nested pairwise comparisons shown in Figure 3.2, comparing the relative effect of lure match (*match*−*mismatch*) on target match/mismatch sentences (*Name* vs. *they*/*I*), and on first-person/third-person intervention sentences (*they* vs. *I*).

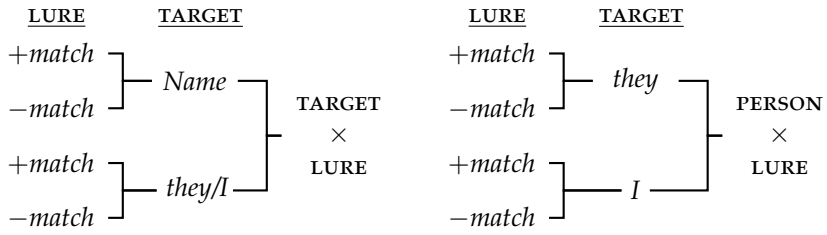


Figure 3.5. Experiment 4b: Nested pairwise comparisons calculated for separate cluster mass permutation tests of TARGET×LURE and PERSON×LURE. Each parent node is the difference (top-bottom) of its daughter nodes

The by-subjects cluster mass analysis revealed a marginal PERSON × LURE interaction at a relatively late time window (2330-2510ms, p=.08), as well as two trending clusters at earlier windows (1140-1270ms, p=0.13; 1310-1440ms, p=.13). The by-items analysis produced similar results, with

a marginal late effect (2290-3320ms, p=0.6), and several earlier clusters which failed to approach significance. Within levels of TARGET, these analyses revealed a prolonged lure-match advantage for progression past reflexives following "they" (by-subjects: 1790-3320ms, p <.05; by-items: 1720-3390ms, p¡.05). The by-subjects analysis also revealed a brief, but significant lure-match *disadvantage* for reflexives following *I*, (290-370ms, p¡.05), but no analogous effect was identified in the by-items analysis. Collapsing levels of target-mismatch, the by-subjects analysis revealed a marginal TARGET×LURE interaction between 610ms and 700ms (p=0.05). However, the by-subject analysis did not reveal any clusters corresponding to a TARGET×LURE interaction.

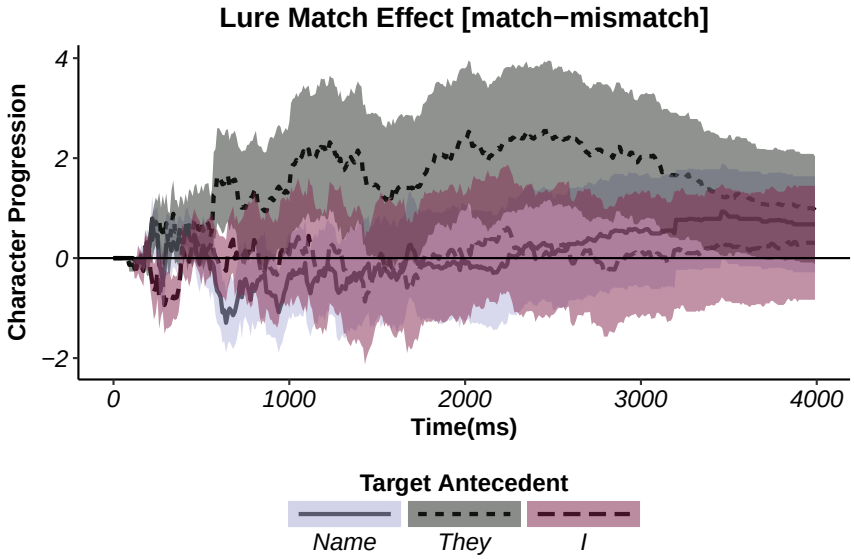


Figure 3.6. Experiment 4b: Mean by-subject lure-match effects (match-mismatch) in characters from first fixation on the reflexive. Error bars represent a bootstrapped 95% confidence interval

These findings generally accord with the results of the fixation duration analyses. To the extent that we observe a significant PERSON×LURE interaction, it occurs in a relatively late time window. This was true in more traditional measures as well, where the interaction only reached significance in go-past reading times at the spillover. In addition, the initial pairwise tests of the cluster mass test showed a significant, sustained lure-match advantage associated with *they* conditions, and no analogous effect for *I* sentences, calling to mind the significant pairwise advantages previously observed in total reading time at the reflexive and spillover regions.

3.3.2.7 Summary

On the whole, Experiment 4b provides perhaps the best evidence so far of person blocking in English reflexive reference. While a statistically significant LURE×PERSON interaction was relegated

to a relatively late reading measure, the overall pattern of behavior is clear, and consistent across several different measures and regions of text. Importantly, these effects obtained despite the fact that *they* is an animate intervener, indicating that indexical pronouns do indeed produce blocking effects above and beyond their status as animate referents might otherwise lead us to expect.

One potentially worrisome observation regarding the results of Experiment 4b, however: unlike in previous eye-tracking studies, lure-match facilitation effects in this study consistently *followed* main effects of target-mismatch. We see a clear effect of target-mismatch in go-past reading times at the reflexive, indicating that participants recognized the violation incurred by a locally-mismatched reflexive. Only after this initial signal—in go-past times at the spillover, and in total time at both regions—do we see lure match facilitation arising in *they* sentences. Characterized in this way, these results seem to indicate a boggle-and-repair strategy employed by readers in this study. This stands somewhat at odds with the results of previous studies, where lure-match facilitation did not seem to be dependent on participants first detecting, and then attempting to repair, a violation of Principle A. While it remains somewhat unclear why this study should be unique in this respect, I think the results are no less interesting if they indicate a repair strategy, rather a default parsing behavior. The fact that English speakers adopt repair strategies grounded in logophoric principles would still, after all, constitute evidence that speakers are (implicitly) *aware* of logophoric principles in the first place, and willing to use them, in a pinch. However, the question of how best to characterize lure-match facilitation (as repair, or first resort parsing strategies) is not the topic of this chapter, and will be revisited in Chapter 5.

3.3.3 Discussion

Much like Experiment 3, Experiment 4 provides telling, if somewhat mixed evidence of person blocking in English. Once again, we find strong lure-match facilitation in off-line judgments and eye-tracking while reading. In measures of sentence acceptability, feature-matched lures substantially improve the acceptability of target-mismatched reflexives. While indexical pronouns reduce the acceptability of sentences overall, Experiment 4a failed to find evidence that they modulate the effect of lure match.

This main effect of PERSON was also found in Experiment 4b, reflecting slower reading times overall for reflexives following indexical pronouns. However, in this case we also saw direct evidence of person blocking, as no evidence of lure-match facilitation was observed in indexical conditions. Importantly, this finding confirms the interpretation assigned to person blocking in Experiment 3. These results corroborate the claim that indexical pronouns inhibit access to non-

local referents above and beyond their status as animate interventionists. Consequently, we have true evidence of *person* blocking, as described by (Huang & Liu, 2001).

3.4 Incorporating perspective into processing models

This chapter has presented repeated evidence that indexical pronouns impede access to lure referents when English speakers are processing and judging reflexive pronouns. The strongest evidence for this claim comes from Experiments 3b and 4b, where a reliable PERSON×LURE interaction was found in both on-line measures of processing difficulty (Experiment 4b), and in off-line judgments of sentence acceptability (Experiment 3b). This effect was also observed in Experiment 3c, albeit with greater variability. Finally, in two acceptability judgments (Experiments 3a and 4a) we observed numerical trends suggesting decreased lure-match facilitation in the presence of indexical pronouns, but failed to observe the critical PERSON×LURE interaction. Collectively, then, these results demonstrate a fairly gradient version of person blocking. Only in Experiment 4b was there a categorical lack of lure-match facilitation in indexical conditions. In all other experiments, lure match facilitation for indexicals was indistinguishable from, or only slightly smaller than, lure match facilitation in non-indexical environments.

Given this mixed picture, the logophlexives hypothesis receives commensurately mixed support. To the extent that we observe interference from indexical pronouns, the basic prediction of the hypothesis seems to be met: referents which introduce more prominent perspective centers decrease concomitantly decrease the salience of lures as perspective holders, reducing their impact on reflexive processing. However, these results are also (at least) surface dissimilar to the description of person blocking given in Section 3.1. Most notably, my characterization of Huang and Liu (2001)'s proposal (working within Kuno (1972)'s notion of direct discourse representation) represented blocking as a grammatical phenomenon in which indexicals *obligatorily* bind the perspective role targeted by logophoric *ziji*. Given this characterization, we might reasonably have expected these effects to be significantly more categorical than they appear.

There are several responses to this concern, ranging from questioning the strength of the original person blocking data in Mandarin, to reconsidering the explanation of person blocking preferred by Huang and Liu (2001). Ultimately, I will argue for an (I think friendly) amendment of Huang and Liu (2001)'s proposal, suggesting that the mechanism behind person blocking is fundamentally non-grammatical in nature. Rather, I suggest it arises from general biases comprehenders bring to bear when determining the perspective center.

3.4.1 Person blocking revisited: considering the data and alternatives

Although person blocking represents a prevalent intuition in the literature, its characterization seems to be somewhat fraught. There is considerable disagreement over exactly how to describe the effect, and what its source may be. Is blocking a categorical, or gradient phenomenon? Does it reflect a hard-coded, linguistic constraint, or does it arise from performance-based factors? In this section I hope to accomplish two things. First, I present evidence that as we observed in our data, person blocking is not entirely categorical. Second, I present two alternative characterizations of person blocking. The first of these represents an attempt to account for person blocking syntactically by taking advantage of agreement constraints (Cole & Wang, 1996). While this approach captures the primary data, we will see that it is empirically inadequate, in addition to being incompatible with the findings of Experiments 3 and 4. The second account is a more semantically sophisticated intellectual relative of Huang and Liu (2001)'s proposal. It presents an attempt at capturing the facts about *ziji* in terms of context shifting operators and logophoric binding (Anand, 2006). The results of this analysis will be more compatible with our findings in this chapter, but will again fail to derive blocking effects from grammatical principles. I turn now to a brief discussion of previous observations of gradience in person blocking.

3.4.1.1 Gradient blocking effects

Experiments 3 and 4 are not the first time gradience has been observed in person blocking phenomenon. In a previous study of person blocking in Mandarin, He and Kaiser (2012) investigated the interpretation of sentence sets like (97). Participants read these sentences in a self-paced reading fashion before answering a comprehension question which probed their interpretation of *ziji* (e.g. *Who can get into a good college?*).

(97) a. Wo gaosu bieren Lisi juede ziji neng kaojin hao daxue.
I tell others Lisi feel SELF able test-in good college
I tell others that Lisi feels like she/I can can get into a good college. (He & Kaiser, 2012)

b. Zhangsan gaosu bieren wo juede ziji neng kaojin hao daxue.
Zhangsan tell others I feel SELF able test-in good college

c. Zhangsan gaosu bieren Lisi juede ziji neng kaojin hao daxue.
Zhangsan tell others Lisi feel SELF able test-in good college

They report two experiments based on this design. In experiment 1, the indexical pronoun used was *wo* (*I*), in experiment 2, it was *ni* (you). In experiment 1, they report fairly weak evidence of person blocking. In fact, they report more third-person non-local responses after a first person

pronoun (97b) than non-local first person responses after third person (97a) (27% vs. 4% matrix responses, respectively). This finding seems to indicate the *opposite* of Huang and Liu (2001)'s reported asymmetry. However, this effect did not replicate in experiment 2, where they found significantly more non-local responses when the local subject was not a second person pronoun (14% vs. 7%). On the basis of these two experiments, the authors conclude that person blocking is *not* a categorical phenomenon, and that second person may be more effective at blocking than first person.

While there are reasons to be concerned with this interpretation[6], the gradient results they report are consistent with the findings of Experiments 3 and 4, suggesting that the effects found in the theoretical literature are not quite so neat as they are reported.

One possible reason for this discrepancy between experimental and theoretical approaches to person blocking lies in the manner in which the influence of non-local referents is tested. In the experiments presented in this dissertation, the effect of "lure match" is alway assessed by comparing responses to sentences in which the matrix subject matches the reflexive against those in which it mismatches. This allows us to derive a sensitivity to non-local referents independent of other manipulations in the sentence. In contrast, at least the reported judgments for Mandarin selectively consider those cases in which the non-local subject could, in principle, act as an antecedent for *ziji*. As a result, it is difficult to distinguish the impact of a local indexical by itself from the interaction of indexicality and non-local reference. In fact, across all five studies presented in this chapter we observed a main effect of PERSON, such that acceptability and reading times were negatively impacted by a local indexical pronoun. In the absence of an independent measure of sensitivity to the non-local referents, then, it remains technically possible that what has been reported as a "person blocking" effect for Mandarin (i.e. an interaction of local person and long-distance reference) is actually merely a main effect of person. More work is needed to demonstrate conclusively that this is not the case.

With that said, I am skeptical that a factorial design approach to *ziji* would reveal only a main effect of person, with no real evidence of blocking. For one thing, local indexical pronouns don't render *ziji unacceptable*, on the whole, they only seem to remove the possibility of long-distance reference. Given this, the analogy to Experiments 3 and 4 here, in which indexicality was perfectly correlated with global unacceptability, is a bit disingenuous. Moreover, the intuitions behind

[6]The fact that non-local interpretations were fairly rare overall suggests that these experiments were not successful in establishing contexts supportive of long-distance binding.

person blocking appear robust, and well attested in the syntactic literature. It would be rash to discard those intuitions as insufficiently rigorous. Finally, and most self-servingly, this chapter has already demonstrated person blocking in *English*—a language with no previous record of it[7]—in the manner prescribed above for Mandarin. Given this, it would, I think, be truly surprising if claims of person blocking in Mandarin were to come to naught.

In sum, there is reason to question the categoricity of the judgments reported in the syntactic literature. Person blocking does not seem to be absolute, nor does it seem to be as ubiquitous as we might have believed. However, far from raising the question of its existence, this observation calls for a closer inspection of how, exactly, it arises. In light of gradience, what are we to make of attempts to grammatically encode blocking behavior? I turn now to two different attempts at tackling this question.

3.4.1.2 Syntactic person blocking accounts

Early accounts of person blocking did not rely on perspective or other discourse properties to derive the effect. Instead, they relied on a series of assumptions about ϕ feature specification in Mandarin, and the syntax of long distance anaphora. I recapitulate here, in brief, the system of Cole and Wang (1996), which is more or less representative of this class of explanation.

First, in explaining *ziji*'s ability to refer outside its governing category, these theories posit a head movement operation which cyclically adjoins *ziji* to successively higher Infl heads[8]. In addition, Cole and Wang (1996) propose that (1) *ziji* and its antecedent must share ϕ features; (2) Infl is featurally underspecified in Mandarin; (3) heads and specifiers must not have conflicting ϕ features; (4) upon adjoining to Infl, *ziji* transmits its features to the head. With this toolkit in hand, the basic blocking pattern can be readily derived. To see how, consider (98), which shows an attempt to long distance-bind *ziji* over a first person pronoun[9]. First, *ziji* moves to adjoin with the embedded I. There, it is assigned a first-person feature, because I, *ziji*, and the embedded subject, *wo*, must all agree. To achieve long distance reference, however, *ziji* must again move, this time the higher I position. Once there, the derivation crashes: *ziji* will transmit its first person feature to the I to which it is adjoined, causing I and its specifier (*Zhangsan*) to have conflicting ϕ features

[7]With one important, notable exception: Loss (2014) reports that at least one dialect of English spoken in North-east Minnesota permits long distance interpretations of embedded reflexives. Moreover, this dialect is also reported to have full-blown person blocking in the same manner as Mandarin. A more careful version of this statement would be that person blocking has not yet been demonstrated for Standard American English.

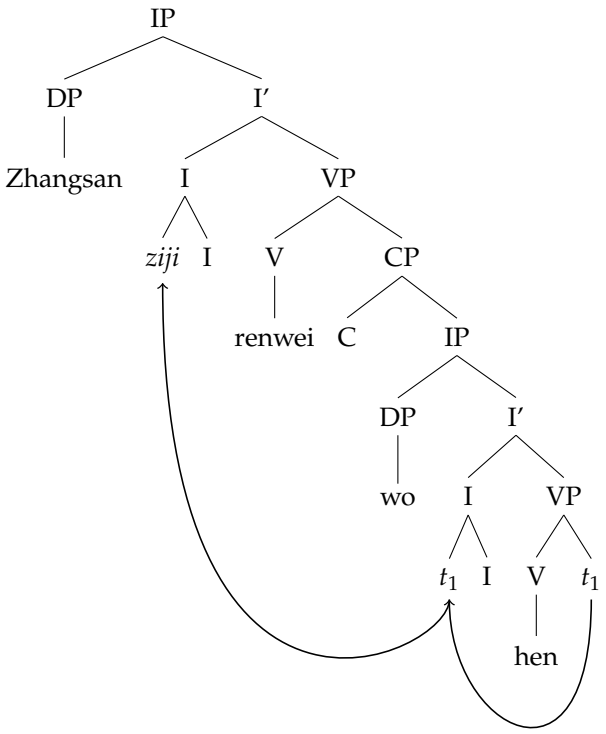
[8]The nature of this movement operation differs among accounts, but these differences will not concern us here

[9]With drastically simplified structure, for convenience.

which, by hypothesis, is disallowed. Conversely, had the embedded subject been a third-person referent, no conflict would have arisen in the aftermath of *ziji*'s movement to the matrix clause, with the result that *ziji* appears to refer long distance.

(98) $*$ Zhangsan renwei wo hen ziji.

Zhangsan$_i$ think I hate SELF$_i$



While this system correctly derives the basic person blocking effect, it fails to account for the other effects described by (Huang & Liu, 2001). First, it incorrectly predicts that third person antecedents should block non-local first and second person. However, as we saw in (91b), reproduced below in (99a), this is not the case. Similarly, this theory has no way of handling the number-asymmetry observed in (93) (reproduced in 99b-c). Given the conflict in number features, local plural features should cause as much of a problem for non-local singular as local singular does for non-local plural. Finally, and perhaps most problematically for this theory, indexicals need not be subjects (or even capable of binding *ziji*) to induce the blocking effect, as seen in (92b), reproduced in (99d).

(99) a. Wo/ni$_i$ danxin Lisi$_j$ hui piping ziji$_{i/j}$.

I/you$_i$ worry Lisi$_j$ will criticize SELF$_{i/j}$

I/you worry that Lisi might criticize me/you/himself. (Huang & Liu, 2001)

b. Lisi$_i$ zidao tamen$_j$ chang piping ziji$_{i/j}$.

Lisi$_i$ know they$_j$ often criticize SELF$_{i/j}$
Lisi knows that they often criticize him/themselves. (Huang & Liu, 2001)

c. Tame$_i$ zidao Lisi$_j$ chang piping ziji$_{*i/j}$.

They$_i$ know Lisi$_j$ often criticize SELF$_{*i/j}$
They know that Lisi often criticize himself (Huang & Liu, 2001)

d. Zhangsan$_i$ gaosu wo$_j$ Lisi$_k$ hen ziji$_{*i/*j/k}$.

Zhangsan$_i$ tell me$_j$ Lisi$_k$ hate SELF$_{*i/*j/k}$
Zhangsan told me that Lisi hates himself. (Huang & Liu, 2001)

As an aside, this hypothesis also seems fundamentally at odds with finding person blocking in a language like English, in which reflexives overtly agree with their antecedents. Given that *himself* and *I* cannot possibly agree, by what mechanism would *I* selectively inhibit access to lure referents? In other words, it is unclear what predictions this theory would make for person blocking in English. Most likely, it would predict simple ungrammaticality in the face of target-mismatch, irrespective of the nature of that mismatch. But this is not what we find. Instead, mismatching some targets seems to produce worse results than others.

In brief, there seems to be little hope for deriving person blocking effects via syntactic agreement mechanisms. Consequently, (Huang & Liu, 2001)'s logophor-based analysis, while not a formal syntactic or semantic model, remains a better explanation of the data. We now turn our attention to a more formal, semantic model to see if it fares better in deriving blocking.

3.4.1.3 Context shifting and logophoric binding

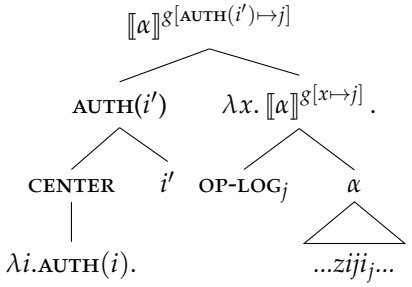
Anand (2006)'s treatment of *ziji* is, in some respects, intellectually related to Huang and Liu (2001)'s, but nonetheless decidedly different. While he agrees with Huang and Liu that *ziji* is behaving (more or less) logophorically, he attempts to capture this behavior with context parameters, rather than movement at LF. Simplifying slightly[10] Anand (2006) actually posits two distinct dialects of Mandarin on the basis of a judgment split among his informants. For one group of speakers, he argues that *ziji* is actually treated as a simple indexical pronoun, the index of which gets overwritten by context-shifting parameters (in his analysis, OP$_{\text{AUTH}}$) optionally embedded under attitude predicates (thus the variability in long-and-local binding). For another group of speakers, he suggests that *ziji* truly is logophoric, and locally bound by a logophoric operator (OP-

[10]Or a lotly.

LOG) which is also (optionally) introduced in the left periphery clauses embedded under attitude predicates. The precise differences between these mechanisms will not greatly concern us here, and so I will focus on the truly logophoric dialect for the present discussion.

Technically, Anand's proposal consists of two pieces: a logophoric operator, and a referential item CENTER, corresponding to a *de se* center (the author of a centered possible world). The logophoric operator carries an index which is mapped onto the variable assignment function of its complement via lambda abstraction, such that the abstracted variable is assigned to OP-LOG's index. By hypothesis, CENTER sits immediately above OP-LOG, with the consequence that CENTER's referent is mapped onto an index in OP-LOG's complement. The semantics of OP-LOG and CENTER are given in (100)[11], along with a schematized structure representing their relation to each other.

(100)



To complete this picture, Anand adopts a version of the system proposed by Koopman and Sportiche (1989), positing a [+LOG] feature associated with *ziji* which must be bound by OP-LOG to be licensed. Ultimately, this produces a system in which *ziji* is *never* bound at distance. Rather, when it receives a non-local interpretation, it is locally bound by LOG-OP, which itself refers to the author picked out by CENTER. Notably, nothing strictly constrains the index complement of CENTER, and so, technically, the *author* referred to by *ziji* is free to vary. Anand prohibits this by stipulating that CENTER's index is bound to the closest available binder.

While Anand's system povides a novel means of talking about logophoricity, he has relatively little to say about the source of person blocking effects[12]. However, he does posit a constraint which derives the effect in his system, given in (101)[13]. This has the effect of disallowing long-distance interpretations past indexical pronouns: if an indexical pronoun is embedded in an attitude predicate, then that particular predicate must not have embedded OP-LOG; therefore there is no operator capable of binding *ziji* and assigning it non-local reference.

---

[11] AUTH(i) is the function which returns the author of a given $<$ *author, addressee, time, world* $>$ tuple.

[12] He is more concerned with capturing the patterns of *de se*/*de re* ascriptions to logophoric *ziji*.

[13] A similar constraint was imposed on shifted indexical *ziji*, stating that *wo* and *ni* cannot be in the scope of the context shifting operator OP_{AUTH}

(101) INDEXICAL POLARITY: *wo* and *ni* cannot be in the scope of OP-LOG

Unfortunately, while this produces the desired effects, it does not clearly derive them. Instead, the constraint appears to codify the fact that *wo* and *ni* must refer to the real-world author and addressee, not an alternative world's. That said, one can clearly see the intuition behind *why* indexicals might avoid OP-LOG: the author supplied by CENTER will not be consistent with the tuple to which the real-world author/speaker belongs. That is, *wo* and *ni* will attempt to refer to the AUTHOR parameter relative to the actual world, but the author parameter has been critically shifted by OP-LOG. However, nothing in the system seems to derive this intuitive conflict.

Thus, while Anand (2006)'s analysis of *ziji* represents a more semantically sophisticated approach, it comes no closer to deriving person blocking than Huang and Liu (2001)'s original proposal. Both accounts express the intuition that conflict over the understood *author* (or SOURCE) in an utterance is response for blocking, but neither fully captures it. Anand's theory stipulates a constraint which accounts for the core person blocking intuitions[14]. Likewise, Huang and Liu provide compelling evidence that blocking is contingent on perspective in some sense, but then note that "some important properties (e.g. blocking effects) of [long distance] reflexive fall outside of syntax in the traditional sense", and, later, that "blocking effects reflect perceptual difficulties that arise when elements within the same discourse domain are 'anchored' to different 'speakers"'.

If this is the characterization they intend, then person blocking seems less and less like a grammatical restriction encoded in speakers' knowledge, and more like a fact about how and when speakers encounter difficulty in tracking perspective in an utterance. That is to say, it's a bottleneck in processing, not a grammatical constraint. Thus, it is related to logophoricity only insomuch as logophoric pronouns seek out perspective centers and indexical pronouns make tracking other perspectives difficult. The final section of this chapter explores this idea.

3.4.2 Person blocking as perspective bias

This chapter began with the premise that person blocking was a grammatical phenomenon related to logophoricity. In the spirit of that premise, Experiments 3 and 4 were aimed at demonstrating these effects for English reflexive reference. However, in the course of collecting and examining the data, this starting premise began to unravel.

First, we saw that Huang and Liu (2001), while giving an intuitive, compelling account of preson blocking, lacked a clear mechanism to derive it in their system. Next, the results of Ex-

[14]Though it isn't clear to me that it would work for the number-blocking asymmetry noted by Huang and Liu (2001)

periments 3 and 4 proved to be more gradient than expected, lending partial support to the logophlexives hypothesis, but raising further questions about the grammatical source of person blocking. Following this concern, we saw that even in Mandarin the effect is more variable than is usually presented in theoretical literature. While reasonably assured of the effect's existence, this prompted a closer examination of some of the alternative models proposed for capturing the facts for *ziji*. In the case of Cole and Wang (1996), the model unfortunately falls short. It neither has the necessary empirical coverage for Mandarin, nor is straightforwardly consistent with finding person blocking effects in languages like English. In Anand (2006)'s system, we found a constraint capable of describing the data, but one which was itself relatively stipulative.

Given the apparent difficulty of deriving person blocking from a grammatical source, I propose that it arises from non-grammatical heuristics, though the nature and representation of these heuristics is still largely speculative. Nevertheless, I suggest that the mechanism which leads comprehenders to preferentially treat speech-verb subjects as perspective centers (as we saw in Experiment 2) is the same mechanism responsible for generating person blocking. Suppose that for a given clause there can only be a single perspective holder (Kuno, 1972; Sells, 1987, i.a.), and that particular linguistic items influence the choice of this perspective center. As we saw in chapter 2, some languages grammaticize these effects, such that perspective is rigidly designate by some particular lexical items. In others, like English, it remains a sub-grammatical tendency, or preference which is expressed by making some perspectives of some entities in the discourse more salient than others. Finally, perhaps person blocking is an indication that some discourse referents (e.g. *speaker*, *addressee*) are *a priori* more salient as perspective holder by virtue of their relation to the current conversation.

Pulling these pieces together, we derive the prediction that logophoric-type reference should be more difficult as more cues accumulate indicating different sources of perspective in the utterance. Working with the data to hand, we should expect, then, to observe a cline in the size of the lure-match facilitation effect as a function of the perspectival salience of the target and lure antecedents. On one end of the scale, inanimate targets (*it*) should produce the greatest degree of lure-match facilitation, as they present no competition for the role of perspective center. On the other end of the scale, target-matched names should produce the smallest about of lure match facilitation: they perfectly match the reflexive and embody all the properties necessary to hold a perspective. In between these two poles, we're left with target-mismatch *they*, and target-mismatch *I/you*. As indexical pronouns, *I* and *you* present perspectives immediately relevant to the current discourse. In contrast, the plural *they* can only hold a perspective insomuch as the group as a whole share

some (literal) viewpoint, and that is the perspective under discussion. Consequently, we should expect more lure-match facilitation following *they* relative to *I*. The full scale (from most facilitation to least) is given in (102).

(102) *It* ≫ *They* ≫ *I/you* ≫ *Name*

Fortunately, given the data collected in Experiments 3 and 4, this is a testable prediction. As the last piece of empirical data in this chapter I present the results of a meta analysis of Experiments 3a, 3b, and 4a, aimed at testing the predictions made by the scale in (102).

3.4.2.1 Judgment data meta analysis

Of the three judgment studies presented in this chapter, only one study produced statistically significant evidence of person blocking (Experiment 3b). However, the effect of lure match facilitation was numerically smaller in the presence of an indexical pronoun in all three studies, suggesting that the effect may exist, but be difficult to detect in off-line ratings tasks. Moreover, in light of the prediction laid out in (102), we have added reason to perform a post-hoc meta analysis analyzing the impact of different target-types on the size of the lure-match facilitation effect.

To accomplish this, data from Experiments 3a, 3b, and 4a were pooled. In this data set, the distinction between *I* and *you* was collapsed into a single level *indexical*. This result in four levels of TARGET: *name*, *indexical*, *they*, and *it*. To test for a gradient effect of target-type on lure-match effects, this factor was reverse helmert-coded into three different contrasts (TARGET: *Name*=1, *Indexical/They/It*=-$\frac{1}{3}$; INDEXICAL: *I*=1, *They*/*It*=-.5; ANIMACY: *They*=1, *It*=-1). The interaction of this factor and sum-coded LURE ($+$*match*=1, $-$*match*=-1) (along with their main effects) were entered as fixed effects in a mixed effects regression fit to the naturalness ratings data from the judgment studies. Random intercepts were included for subjects and items, but random slopes were only fit to these groups for the effect of LURE. Since not all levels of TARGET were represented for every time, and not every subject responded to every level of TARGET, it would have been inappropriate to include random slopes for levels of target match. All rejection criteria that were applied to experiments 3 and 4 were applied to this meta analysis.

RESULTS By-subject mean ratings in the meta-analysis are given in Table 3.6. In addition, this table presents the model-fit results of a model which evaluated pair-wise comparisons of lure-match within each level of TARGET. The mixed effects model fit to responses revealed significant main effects of TARGET, INDEXICAL, and LURE, indicating respectively: increased acceptability for target-matched reflexives ($\hat{\beta}$=1.48, t=45.48); decreased acceptability associated with indexicals ($\hat{\beta}$=-

0.23, t=6.85), and increased acceptability for feature-matched lures ($\hat{\beta}$=0.18, t=8.27). These main effects were qualified by three significant interactions. First, there was a significant TARGET×LURE interaction ($\hat{\beta}$=-0.21, t=6.71), indicating that *name* conditions were associated with substantially less lure-match facilitation than the remaining conditions. Second, we observed a significant INDEXICAL×LURE interaction ($\hat{\beta}$=-0.08, t=2.51), indicating a smaller lure-match effect for indexical sentences relative to *they* and *it* sentences. Finally, there was a significant ANIMACY×LURE interaction, indicating greater lure-match facilitation when the target was inanimate, than when it was animate. A full table of fixed effects is given in Table A.11, in the appendix.

Table 3.6. Meta Analysis: Mean by-subject naturalness ratings (standard error in parentheses) for the meta analysis of Experiments 3a, 3b, and 4a, collapsing across first and second person pronouns ("Indexical"). Lure match effects represent pairwise comparisons of lure match nested within levels of target.

	Lure		Lure Match Effect	
Target	*+match*	*−match*	$\hat{\beta}$	t
It	3.93 (0.12)	3.14 (0.11)	0.39 (0.04)	8.85
They	3.57 (0.09)	3.16 (0.10)	0.20 (0.06)	3.55
Indexical	3.29 (0.11)	2.95 (0.11)	0.17 (0.04)	4.89
Name	5.29 (0.08)	5.37 (0.07)	-0.04 (0.04)	1.11

The results of this analysis directly confirm the predictions given in (102). When results are pooled across studies, we see that labeling the effects observed in separate experiments "person blocking" is misleading: blocking occurs in several measures as a function of how well a referent (1) matches the reflexive (target match/mismatch contrast), and (2) how salient the target is a perspective center. Person features play no particularly special role, except inasmuch as they denote referents whose perspective is highly salient for the current discourse.

3.4.3 Wrapping up blocking effects

In this chapter, I presented the results of two experiments designed to find evidence of person blocking in English. The goal of these experiments was to show that English, like Mandarin, is sensitive to local-discourse participants when assigning non-local interpretations of reflexives. This outcome was achieved: indexical pronouns in fact impeded access to lure referents, albeit in a gradient, and often transitory manner. On the whole, then, this chapter provides continued support of the logophoric reflexives hypothesis. Comprehenders attend to lures which act as logophoric antecedents. Lures are less likely to act as logophoric antecedents when a local discourse participant intervenes because this participant renders the lure a generally less salient perspective holder. In this sense, this chapter has presented the converse of chapter 2. In that chapter, we saw how manipulating the salience of the lure made it more or less available for an embedded

reflexive, whereas in this chapter, we saw that manipulating the salience of the target impacts the relative accessibility of the lure.

As a consequence of this investigation, we have also come to a refined understanding of person-blocking. The greater portion of the recent discussion has been spent arguing that person blocking should not be understood as a grammatically encoded phenomenon, but rather as the result of the heuristics used to locate a perspective center in the discourse. While I have had to remain vague about the particulars of these heuristics, the meta analysis of the previous section demonstrated that blocking should not be considered a categorical phenomenon predicated on person. Instead, it seems to be a gradient function of how salient a given referent is in the local discourse.

Overall, the conclusions of this chapter resonate strongly with those reached at the end of chapter 2. While some languages may grammatically encode perspective with linguistic structure, many others co-opt linguistic cues as a means of gradiently tracking perspective.

CHAPTER 4
ANIMACY IN REFLEXIVE PROCESSING

It is a fundamental property of logophoric pronouns that they must take an animate antecedent. This follows naturally from the fact that logophors refer to the perspective holder of an utterance, a role which must be filled with a consciousness center. Since inanimate referents cannot act as consciousness centers, they cannot act as the perspective holder, and therefore cannot antecede a logophoric pronoun. From this, it follows that only animate reflexives should be able to be used logophorically—an inanimate reflexive would require an inanimate, and therefore inherently non-logophoric, antecedent. In previous work, Charnavel and Sportiche (2016) leveraged this observation to test the long-distance binding potential of anaphors in French. As discussed in Chapter 1, they report that only animate reflexives may take non-local antecedents in French, while inanimate anaphors are always obligatorily locally bound. This led them to suggest that long-distance interpretations of anaphors in French are logophoric.

Animacy thus serves as a useful diagnostic for logophoric behavior. To the extent that logophoricity drives lure-match effects in reflexive comprehension, we should expect that these effects should be contingent on the animacy of the reflexive and its potential referents. More precisely, if lure-match facilitation in previous studies was due to logophoricity, then these effects should not extend to inanimate reflexives. To date, no psycholinguistic study has investigated inanimate reflexive pronouns, much less contrasted their behavior with their animate counterparts. In this chapter, I present an acceptability judgment study aimed at doing precisely this. Experiment 5 uses the mismatch paradigm to directly compare the behavior of animate and inanimate reflexive pronouns. If lure-match facilitation is the product of logophoricity, then inanimate reflexives should be insensitive to lure referents.

4.1 Experiment 5: Animacy in reflexive comprehension

To assess the impact of reflexive animacy on sensitivity to lure referents, sentences were manipulated as in (103). All items were bi-clausal, with a reflexive in the embedded direct object position. This reflexive was manipulated so that was either animate (*him/herself*), or inanimate

(*itself*). As in previous studies, the embedded and matrix subjects were manipulated so that they either matched, or mismatched the embedded reflexive. Reflexive mismatch was realized as a mismatch in animacy (eg. animate referents paired with an inanimate reflexive).

(103) The {budding author / press release} wrote that the {brilliant detective / amateur detectives union} credited {himself / itself} with exposing the hidden crime syndicate.

Since inanimate referents cannot act as logophoric antecedents (they do not represent a consciousness center), the logophlexives hypothesis predicts that animate reflexives should be more sensitive to the manipulation of the matrix subject than inanimate reflexives. Thus, in sentence rating and reading studies, we should observe a stronger lure-match effect for animate reflexives than their inanimate counterparts. To test this prediction, materials patterned on (103) were included in a sentence acceptability survey. An eye-tracking study investigating these items is currently underway, but there is currently insufficient data for analysis. Thus, only the results of the acceptability judgment study are reported here.

4.1.1 Experiment 5: Acceptability judgments

64 self-reporting native English speakers were recruited via Amazon Mechanical Turk and compensated $4 for their participation. Prior to analysis, ten participants were excluded for reporting exposure to an East Asian or West African language, on the basis of age (participants older than 55), or for participating in a prior experiment about reflexive pronouns. The remaining 54 participants were between the ages of 21 and 55 (median age: 31). 25 of these participants identified as male, and 29 as female. Participants were more or less equally distributed across 22 different states, with the exception of California and Florida, which were the home-states of 7 and 5 participants, respectively.

4.1.1.1 Materials

48 items patterned on (103) were created and interleaved with 72 sentences from unrelated experiments in a Latin square design. Of these filler items, 48 represented an agreement attraction paradigm, introducing grammatical errors in agreement into the experiment. The remaining 24 items were all grammatical, and contained indirect questions and embedded WH extraction. In total, 40% of the items in the experiment were normatively ungrammatical.

4.1.1.2 Procedure

The experiment was coded and hosted online using the Ibex Farm[1] software for web-based experiments. Participants were instructed to rate sentences on a scale from 1 (*very unnatural*) to 7 (*very natural*), and given four sentences exemplifying the end points of the scale (two each) as practice. Sentences were presented above the scale, with the endpoints labeled "completely unnatural" and "completely natural". There was no time limit on responses. Participants indicated their rating by either clicking the on-screen number, or pressing the corresponding number key. The experiment lasted approximately 45 minutes.

4.1.1.3 Analysis

Sentence ratings were analyzed with linear mixed effects regression, taking the factors LURE (+*match*=1, −*match*=-1), TARGET (+*match*=1, −*match*=-1), REFLEXIVE (*animate*=1 vs. *inanimate*=-1), and all interactions as fixed effects. Random slopes and intercepts were estimated for both subjects and items, though correlations between the random effects were excluded from the model. Planned pairwise comparisons were evaluated by nesting the factor LURE inside the factors TARGET and REFLEXIVE, testing for an effect of lure match within each target-match/reflexive-type pair. *t*-values of absolute value $\geq$2 were taken to be significant (Gelman & Hill, 2007). To account for inordinately long (or short) response latencies (indicating either lack of attention, or accidental button presses, respectively) response times were z-score transformed by subject prior to analysis. Trials with $|z| > 3$ were then rejected, resulting in the exclusion of 2.4% from analysis.

4.1.1.4 Results

By-subject mean ratings are given in Table 4.1. Mixed effects modeling revealed a substantial main effect of target match, reflecting the fact that participants rated target mismatch sentences significantly worse than their target match counterparts ($\hat{\beta}$=1.08, t=11.01). No other effects reached significance, although there were trending TARGET×REFLEXIVE ($\hat{\beta}$=0.07, t=1.76) and TARGET×LURE×REFLEXIVE ($\hat{\beta}$=−0.04, t=1.43) interactions. The TARGET×REFLEXIVE trend is likely driven by the fact that target-matched animate reflexives were rated slightly better than inanimates, while target-mismatched inanimates were rated slightly better than animates. The trending three-way interaction possibly corresponds to the numerical trend by which LURE exerted the largest influence (δ=.22) on target-mismatched, animate reflexives. However, while this pairwise

[1] http://www.spellout.net/ibexfarm

difference trended significant in the nested model ($\hat{\beta}$=0.11, t=1.75), it failed to reach significance. Notably, no other pairwise comparison of the lure match effect approached significance (all t<0.7).

Table 4.1. Experiment 5: Mean by-subject naturalness ratings (standard error in parentheses)

		Reflexive Type	
Target	**Lure**	*Animate*	*inanimate*
+match	*+match*	5.7 (0.11)	5.64 (0.10)
	−match	5.78 (0.12)	5.56 (0.12)
−match	*+match*	3.55 (0.16)	3.62 (0.17)
	−match	3.33 (0.17)	3.58 (0.18)

4.1.2 Summary

Like Experiment 1a, Experiment 5 revealed robust effects of target match, and little to no effect of lure match on off-line judgments. To the extent that we observed any influence of lure match in this study, it was confined to the animate reflexive, target-mismatch conditions. However, this effect was quite small, and failed to reach significance. Moreover, the REFLEXIVE×TARGET×LURE interaction did not reach significance, despite trending in the expected direction. Consequently, no strong conclusions about the conditions under which lure-match effects are observed may be drawn from this data. The evidence weakly suggests that animacy may be a prerequisite for long-distance reference, but does not directly support this position. In this respect, the findings of Experiment 5 are very similar to those observed in Experiment 1a, a parallel which is picked up in the general discussion, below.

4.2 Discussion

Experiment 5 attempted to find evidence that animate, and inanimate reflexives were differentially prone to lure-match facilitation, as predicted by the logophlexives hypothesis. While the results of this experiment are suggestive (numerically trending in the expected direction), they do not overtly confirm these predictions. Instead, they seem to mirror the effects of Experiment 1a: strong target-mismatch effects with little to no effect of lure-match facilitation. In contrast, Experiments 2-4 all found substantial lure match facilitation effects in off-line measures of sentence acceptability. This pattern is notable given that, in many respects, the manipulations in Experiment 5 most closely parallel those in Experiment 1. These experiments differed from Experiments 2-4 on two critical dimensions: (*i*) Experiments 1 and 5 did not include context sentences; (*ii*) Experiments 2-4 used pronominal forms for the embedded subject in target-mismatch conditions, while Experiments 1 and 5 used full DPs in all conditions. It may be that these factors contribute to overall smaller degrees of lure-match facilitation in off-line judgments.

With respect to the first point, Experiments 2-4 all included context sentences preceding the critical target sentence. In contrast, experiments 1 and 5 presented the critical experimental sentences in isolation, without the benefit of context. Perhaps perspective taking is not so easily accomplished in the absence of a rich discourse context, in which case logophoric reference may be less sustainable without adequate contextual support. While the contexts of Experiments 2-4 were fairly minimal, it may be that they were sufficient to support (some degree of) logophoric reference and give rise to lure-match effects capable of surviving in off-line judgments.

Alternatively, the critical difference among judgment studies may lie in the form of the embedded subject. In all of the judgment studies in which a strong lure-match facilitation effect was observed, the embedded subject was a pronominal, rather than a full DP. Indeed, in judgments, at least, the strongest "blocking" effect we observe isn't with indexical pronouns, but rather with full DP local referents. At present, the causal role played by this factor is somewhat unclear. It may be that pronouns are simply less likely to control the perspective center than full DPs. However, this explanation seems unsatisfying given (*i*) the strong lure-match effects observed in on-line measures of processing difficulty in Experiment 1b, and (*ii*) the cross-linguistically attested phenomenon of person blocking (which necessarily occurs with pronominal referents). Alternatively, pronouns may influence the implicit prosody associated with the embedded direct object position. Since pronouns are generally less likely to receive primary stress (Selkirk, ????), primary stress is more likely to be associated with another position. In this case, that alternative position may well be the embedded direct object. As noted by Ahn (2015), local interpretations of reflexives are canonically realized with the reflexive remaining unstressed. Taking these observations together, an embedded subject pronoun may lead to an embedded, direct object reflexive bearing primary stress, thereby decreasing its association with local-binding and increasing the likelihood of taking a long-distance antecedent. If this account is correct, then full DP subjects are more likely to assist in enforcing local binding, thereby explaining the relatively small effect of lures in Experiments 1a and 5.

In brief, there are multiple possible sources for the small effects observed in this, and previous experiments. These range from the lack of context sentences, to the possible role played by pronominal subjects. Regardless of its source, the lack of effect in Experiment 5 should be treated with some caution. Beyond the explanations offered above, this experiment only includes off-line judgment data, while most of the critical effects in Experiments 1-4 were observed in eye-tracking while reading. Given this, it may be that a reliable difference between animate, and inanimate

reflexives will emerge in more sensitive measures like eye-tracking while reading. At present, however, we have no concrete evidence of this difference.

CHAPTER 5
LOGOPHORICITY AND REFLEXIVE COMPREHENSION

In Section 5.1, I review the primary findings of Chapters 2-4 and make the case that Principle A fallibility in sentence comprehension should be seen as an expression of logophoricity. Following this argument, we examine more closely the time-course of the effects we have been observing. The experiments in preceding chapters have shown considerable variability in the timing of the critical effects, both in terms of which particular measures and regions exhibit effects, and in terms of whether they survive in off-line measures like acceptability judgments. With this executive summary of the findings in hand, Section 5.2 presents a model which incorporates logophoricity into existing accounts of reflexive antecedent identification. In brief, this model proposes that a silent logophoric operator (OP_{log}) acts as a local binder for reflexives, and that it is the referential properties of this operator which give rise to the effects discussed here. From there, sections 5.3 and 5.4 consider the implications of this model for our understanding of retrieval models and Binding Theory respectively. The chapter then concludes with a look towards directions for future inquiry.

5.1 The case for a logophoric source of Principle A fallibility

This dissertation began with an examination of various approaches to Binding Theory, broadly labeled the "standard" model and "predicate-based" theories. While these theories make the same predictions for argument reflexives (they should be obligatorily bound), they differ in their treatment of non-argument reflexives. Under the standard model, this distinction among syntactic positions is not relevant, as reflexives in every position are expected to be subject to a local binding constraint (Chomsky, 1986; Charnavel & Sportiche, 2016). However, predicate-based models lead us to expect that non-argument positions should be exempt from obligatory local-binding, and should instead be able to find their antecedent from a discourse model (Pollard & Sag, 1992; Reinhart & Reuland, 1993). Importantly, these "exempt anaphors" seem to be constrained by alternative factors, and need to refer to the perspective holder of an utterance, leading proponents of these models to label such uses "logophoric" (Reinhart & Reuland, 1993).

However, when the predictions of these models are compared to findings in the sentence processing literature, a double-dissociation of theoretical predictions and real-time behavior is observed. First, non-argument reflexives do not always behave like exempt anaphors in early stages of sentence comprehension. Cunnings and Sturt (2014) found no evidence that comprehenders attend to non-local antecedents even when the reflexive is in a non-argument position (e.g. embedded in a possessorless NP). Conversely, Kaiser et al. (2009) present evidence that the presence of a co-argument does not always constrain interpretation to local antecedents. In their study, participants chose subject antecedents for object-NP-embedded reflexives even when this NP had a possessor. Finally, a strict, early application of any of these theories is challenged by the results in Parker and Phillips (2017), whose studies report lure-match facilitation associated with direct object reflexives.

The double dissociation in behavior notwithstanding, these findings do not, necessarily, challenge syntactic theories of binding. It is possible that these effects arise entirely as artifacts of the sentence processing mechanism. Under this view, Binding Theory (pick a flavor) accurately describes a speakers' knowledge state, but this knowledge is imperfectly implemented for real-time deployment. This seems to be the preferred interpretation in the sentence processing literature, where appeals to a noisy, error-prone, retrieval mechanisms are used to explain "grammatical fallibility". However, predicate-based models *have* faced more recent challenges in the syntactic literature. Charnavel and Sportiche (2016) note that inanimate reflexives in French behave unexceptionally like obligatorily locally bound anaphors. Only animate reflexives seem to be capable of finding non-local antecedents. Given this, they suggest that there is no principled division between argument and non-argument reflexives, but instead that non-local reference arises from animate reflexives being bound by a logophoric operator, which itself refers outside the local domain. In particular, this logophoric operator (and, consequently, the bound reflexive) refers to the perspective center of the utterance in the same way as logophoric pronouns in other languages (Sells, 1987; Culy, 1997).

This alternative view of non-local reflexive reference raised an interesting possibility: perhaps variability in the sentence comprehension literature was similarly tied to logophoric binding. If so, we would expect sensitivity to lures to be conditioned on their viability as logophoric antecedents. The "logophlexives" hypothesis, then, has been the central driving force of the investigations presented here, with the goal of showing that sensitivity to lure referents patterns with constraints known to hold of logophoric pronouns cross-linguistically. I have presented five experiments aimed at doing so, and summarize their primary findings below.

5.1.1 Primary findings and arguments

In Chapter 1, it was shown that languages vary in the kinds of predicates capable of embedding logophoric reference, leading Experiments 1 and 2 to investigate the impact of attitude verbs on reflexive comprehension in English. Based on the discussions in Sells (1987) and Culy (1997), logophoric reference is expected to be easiest when the logophor is embedded under a speech predicate, and relatively difficult when it is embedded under a perception predicate. This is because every instance of a logophor may refer to a SOURCE role (in Sells' terms) or licensed by a speech verb (to use Culy's system). The logophlexives hypothesis, then, likewise predicted greater sensitivity to lures which were the subjects of speech verbs, relative to those which were the subjects of perception verbs. Broadly speaking, this prediction was confirmed.

In Experiment 1, sensitivity to the features of lure referents was shown to be dependent on two factors: (*i*) whether the target antecedent matched the reflexive, and (*ii*) the kind of verb used to embed the reflexive pronoun. In general, feature-matched lures facilitated reading times for target-mismatched reflexives if those lures were the subjects of speech verbs, but not perception verbs. However, in Experiment 2 this distinction between verb types disappeared when the lure was the only consciousness center in the utterance. That is, when there was no ambiguity in which referent holds perspective, the type of attitude verb used to embed the reflexive pronoun no longer mattered. Based on these findings, I suggested that attitude verbs do not strictly grammatically constrain reference to their subjects, but rather probabilistically affect the likelihood with which their subjects will be taken as the perspective center of an utterance. When a sentence contains multiple consciousness centers, this factor helps the comprehender locate the relevant perspective center. However, when only one consciousness center is available, this factor is redundant, and does not (immediately) impact perspective tracking. Thus, embedding verbs can impact lure-match facilitation, but only in the face of ambiguity in the intended perspective center.

These results are consistent with the theories proposed by Sells (1987) and Culy (1997), albeit more gradiently than either of these authors envisioned. In particular, they suggest that logophoric reference (in English, at least) does *require* particular attitude verbs to be licensed. However, constructional (i.e. non-lexical) licensing was already necessary to account for logophoric uses of pronouns in Japanese (*zibun*) and Ewe (*yè*). In these languages, logophoric pronouns may occur in causal clauses, un-embedded under any attitude verb. If we allow that the perspective center is always constructionally assigned in English, and that this assignment is a probablistic function of the local context, then we can allow verb-type to be an influential factor in perspective taking. Furthermore, this suggests a model in which English exists on the probablistic end of pressures

which, eventually, are grammaticized in languages like Ewe. This, and related questions of language evolution will be revisited in the conclusion. For the present, note that English reflexive processing *is* sensitive to verb classes in a manner consistent with the behavior of logophoric pronouns, albeit not categorically so.

Continuing the theme of perspective taking as the primary goal of logophoric reference, Experiments 3 and 4 investigated situations in which the perspective-setting properties of the target antecedent influence access to the lure referent. Of interest here was the observation that the Mandarin reflexive *ziji* demonstrates a "person blocking" effect when assigned a non-local interpretation (Cole & Wang, 1996). This effect was argued to result from the logophoric nature of long-distance *ziji* (Huang & Liu, 2001; Anand, 2006). In this effect, the presence of an intervening indexical (first or second person) pronoun prevents *ziji* from taking a non-local antecedent (that isn't the indexical pronoun, itself). Following Huang and Liu (2001) (and, more abstractly, Anand (2006)), I suggested that person blocking arises from competition between indexical pronouns and other referents for the role of perspective holder. Critically, indexical pronouns are posited to hold special salience for this role due to their status as local participants in the discourse. Given this, Experiments 3 and 4 explored the possibility of person blocking for English reflexive reference.

Taken together, these experiments provided fairly consistent evidence of person blocking. In Experiment 3, there was a repeated numerical trend consistent with diminished sensitivity to lures in the presence of indexical pronouns. This effect reached significance in one off-line judgment study (Experiment 3b), and in two measures of reading difficulty in Experiment 3c. Experiment 4 replicated these effects, demonstrating the same numerical trend in judgments (Experiment 4a), and a reliable effect on reading times in Experiment 4b. Importantly, Experiment 4 demonstrated that the effect of person was *not* a contrast in animacy, as the foil target-mismatch condition involved an animate mismatch in number. Regardless, across all of these studies, there was a consistent decrease in sensitivity to lure referents in the presence of an indexical pronoun. As in Experiments 1 and 2, however, this effect was gradient. Indexical pronouns did not *categorically* inhibit reference to the matrix subject, they merely impeded it. While this gradience is at odds with the categorical effects reported in the syntactic literature (Cole & Wang, 1996; Huang & Liu, 2001; Anand, 2006), they are consistent with experimental investigations of person blocking in long-distance sues of *ziji*, which find a non-categorical influence of indexical pronouns (He & Kaiser, 2012). Moreover, all five studies presented in these experiments demonstrated a consistent main effect of person, indicating a stronger target-mismatch penalty for mismatch in person, rather than number and/or gender.

As with Experiments 1 and 2, Experiments 3 and 4 suggest that effects which appear categorical in the literature on logophoricity surface gradiently in English. There are two logically possible solutions for this disparity. First, it may be that the stringency of constraints on logophoric pronouns has been misrepresented, and that speakers are more permissive in their uses than is currently reported. This seems to be the case for person blocking, at least, where experimental evidence (He & Kaiser, 2012) finds fairly gradient effects in Mandarin, and the theoretical literature itself contains disagreements over which configurations induce blocking (Cole & Wang, 1996; Huang & Liu, 2001; Anand, 2006). Given this, we might question whether constraints on embedding verbs as reported in (Culy, 1997) are likewise more porous than previously believed. Adopting this position would, in effect, make behavior in logophoric languages more similar to the behavior observed here for English.

Alternatively, it could be the case that constraints on logophoric pronouns emerge from general cognitive pressures associated with identifying and tracking perspective holders. If so, it seems reasonable to expect a continuum of behavior, from languages which are probabilistically sensitive to these pressures (e.g. English) to languages which have fully grammaticized it (e.g. Ewe). The question then is why reflexive pronouns, in particular, seem to be sensitive to these pressures in ways that other referring devices are not. In other words, what's the connection between logophoricity and reflexivity that leads to the repeated merging of the two forms across the world's languages? As will be seen in Section 5.2, this relationship emerges from the fact that logophoric reference is simply a special case of local-binding for reflexives. Deeper connections between logophoricity and reflexivity will be explored in greater detail in the conclusion of this work.

Finally, Experiment 5 adopted Charnavel and Sportiche (2016)'s insight and tested for differences between animate, and inanimate reflexives in English. Critically, inanimate reflexives cannot take an animate antecedent, and therefore cannot act as perspective-sensitive elements—there is no consciousness center to which they can refer. Therefore, if Principle A fallibility is conditioned on a logophoric interpretation, then we expected to find no lure-match facilitation for inanimate reflexives. Unfortunately, this experiment provides little, if any evidence of logophoricity. No significant lure-match facilitation effects were observed, in contrast to preceding studies. It is at present unclear why this experiment failed where the others succeeded, and more work is needed to understand the puzzling lack of effects observed.

Nevertheless, four of the five experiments (consisting of 10 studies, in total) presented in this dissertation have provided consistent evidence that Principle A fallibility in sentence comprehension is tied to constraints on logophoric interpretations. English speakers are willing to entertain a

lure antecedent for a reflexive when that lure represents the perspective from which an utterance is reported, and is therefore a good logophoric antecedent. These findings challenge the view that Principle A incompatible behavior truly represent "fallibility" in the online implementation of grammatical knowledge. That is, the results of seen in Chapters 2-4 don't seem to be the product of *errors* made by in a noisy memory search process, but rather arise from grammatical principles which exist alongside Binding Theoretic constraints. This argument against an "errorful" interpretation of logophoric effects is made more forcefully in the Section 5.3.2 in an extended comparison with various "grammatical illusions". In the mean time, we turn to a closer consideration of the time course of these effects.

5.1.2 The timecourse of logophoricity

Across Experiments 1-5, the time course of critical effects was fairly varied. In most, but not all, cases, we saw substantial lure-match effects both in on-line (eye-tracking while reading), and off-line (acceptability judgment) measures of processing difficulty. However, within each category, effects were not evenly distributed. In on-line measures, we observed quite early effects in two of our experiments, and relatively late effects in the remaining two. In off-line measures, we found substantial lure-match effects in four out of five judgment studies[1]. This variability deserves attention, and that is the purpose of this section. I address the differences between eye-tracking studies first.

In Experiments 1b and 3b, we saw lure-match facilitation effects in early measures of reading difficulty (first pass and go-past reading time) at the critical reflexive region. Critically, these effects of lure were modulated both by verb type, and by intervening indexicals. On the basis of these studies, one might reasonably conclude that lures impacted reflexive comprehension quite rapidly, and that this impact was modulated by logophoric principles. However, in Experiments 2b and 3b, the critical effects of lure (and associated logophoric modulations) arose in the spillover region, primarily in go-past and total reading time—markedly later than the effects observed in Experiments 1b and 3c. Given this, one might reasonably wonder whether the effects reported here reflect default processing operations, or later repair operations (Lago, Shalom, Sigman, Lau, & Phillips, 2015).

In response to this ambiguity, I suggest first that the effects reported here are no less interesting if they represent repair strategies rather than first-pass procedures in sentence comprehension. If

[1] Results of Experiment 5 pending.

the effects in Experiments 1-5 *do* represent a repair strategy, then they exhibit behavior which is (*i*) inconsistent with the surface grammar of English, (*ii*) consistent with grammatical principles in other languages. In other words, such behavior would start to look like the "emergence of the unmarked" often discussed in phonology: default, cross-linguistically unmarked behavior emerges when normally preferred grammatical constructions are unavailable. Inasmuch as this indicates a universal inventory of preferred linguistic behavior, this constitutes a fairly remarkable finding. This is an idea to which we will return in the conclusion of the present work.

However, there remains fairly compelling evidence that the effects observed here are not the product of some post-hoc repair strategy recruited to fix a perceived error. First, there's the fact that lure-match facilitation (and concomitant logophoric constraints) emerge quite early in Experiments 1b and 2c. In these studies, the critical effects appear to be part of first-pass processing of the unfolding sentence. The relatively late effects in Experiments 2b and 4b might then be attributed to cross-experiment (and cross-participant) variability, rather than indicating a systematically late effect[2]. In general, effects of syntactic processing seem to be more variable in their time course than, for example, lexical processing (Clifton, Staub, & Rayner, 2007). Moreover, Experiment 1c speaks against this interpretation. In Experiment 1c, we saw that participants adopted non-local interpretations of reflexives as much as $\frac{1}{3}$ of the time, even though the target matched the reflexive. Consequently, non-local interpretations cannot simply be the result of a repair operation, as no repair was necessary in these cases. Finally, with attentive observation, it is possible to find examples of logophoric uses of reflexives in naturally occurring speech, as seen in (104). The existence of such examples is strong evidence against a repair-type strategy, as when the speaker produces such uses of reflexives, no error has yet been made. Given this, an analysis of logophlexivity as error-repair seems untenable.

(104) a. It wasn't until too late that Michael$_i$ realized that *hermano* was Spanish for *brother*, and that the person Marta was infatuated with was, in fact himself$_i$.

(Arrested Development, S01E13: *Beef Consomme*)

b. Our president$_i$ can only understand the world to the extent that it involves himself$_i$.

(Last Week Tonight with John Oliver, S04E13: *Stupid Watergate*)

[2]Interestingly, all three experiments in which the target was a pronoun (Experiments 2b, 3c, and 4b) had higher base-rates of reflexive skipping (20%) relative to the experiment in which the target was a full DP (Experment 1b, 15% skipping rate). This may explain some of the variability in the experiments, with effects at the spillover region observed in those studies in which more reflexive skipping occurred.

c. I know exactly what's wrong with it, I just don't know how to fix it. It's like a metaphor for myself. (Elaine Teng, p.c.)

Turning now to off-line measures, in four of five judgment studies we observed substantial lure match effects. In Experiments 2a, 3a, 3b, and 4a, there were sizable main effects of LURE, such that feature-matched lures improved sentence ratings. Moreover, there were significant TARGET×LURE interactions in all five studies, indicating that lures exerted a greater influence on target-mismatch reflexives than on target-matched reflexives. Despite this consistency, only one study found a statistically significant critical interaction (Experiment 2b: PERSON×LURE), though the remainder showed numerical trends in the expected directions. Collectively, these effects indicate that, in off-line judgments of acceptability, participants are more susceptible to morphosyntactic match than they are to logophoric constraints. That said, logophoric constraints still seem to exert an influence: the perspective taking capabilities of the lure and target impacted the judgment process. The one case in which we failed to observe a main effect of LURE (Exeriment 1a) was also the only judgment study with a fully referential DP in the target position. Thus, lures seemed to exert the least influence when the target was a fully referential DP capable of holding a perspective. In contrast, as we saw in the meta analysis in Chapter 3, the influence of lures on acceptability steadily increases as the target becomes less salient, and the lure more salient, as a perspective holder. Given this, lure-match effects and logophoric constraints seem both to be well represented in off-line measures of sentence acceptability. In other words, both morphosyntactic, and logophoric constraints seem to be fairly stable in off-line measures of sentence acceptability.

5.1.3 On multiple match effects

There is one final component of the data which bears discussion: the repeated failure to find an effect of lure-match on target-matched reflexives. Recall that in Badecker and Straub (2002) post reflexive material was read more slowly when it was matched by more than one referent. This effect is expected under most basic cue-based retrieval models (Lewis & Vasishth, 2005; McElree, 2000; Van Dyke & McElree, 2006), though more recent instantiations eschew this prediction (Jäger et al., 2017). Regardless, we see no evidence across all eleven studies in Experiments 1-5 that feature-matched lures increase the difficulty of processing target-matched reflexives. We *do* see that participants are sensitive to lures, even in target match environments (Experiment 1c), but we see no evidence that this actually engenders greater difficulty in reflexive comprehension. Consequently, the existence of multiple-match effects, and the viability of the models which predict them, seems to be in question (see also (Jäger et al., 2017) who reach a similar conclusion).

One possible reason for which no multiple-match effects were found lies in the ambiguity advantage effect (Traxler, Pickering, & Clifton, 1998; Swets, Desmet, Clifton, & Ferreira, 2008, i.a.). In this case, structures which are globally ambiguous are processed more easily than their disambiguated counterparts. This effect has recently been shown to generalize to pronominal reference, indicating that globally ambiguous pronouns are read more quickly than their locally disambiguated counterparts (Grant, Dillon, & Sloggett, 2015). It may be that a similar effect obtains in the multiple-match conditions for reflexives, such that multiply-matched reflexives are considered "ambiguous", and read generally quickly. Alternatively, it may be that locally-matched referents present such tempting antecedents that the matrix subject rarely, if ever, is able to provide viable competition. This is closer to the tack taken in (5.2.2). In the meantime, there seems to be good reason to doubt the existence of multiple-match effects in reflexive comprehension: logophoric (lure-match) effects only seem to surface in the face of target-mismatch, an effect labeled here the "target-match" asymmetry. In discussions of plausibly related phenomena (e.g. agreement attraction), this kind of effect is often called the "grammaticality asymmetry" (Wagers et al., 2009, i.a.). Given that I am arguing that logophoric reference is not *ungrammatical*, I adopt the more neutral "target-match" terminology here.

5.2 A model of logophlexivity

With these facts in hand, we attend now to the question of how best to model logophoric behavior in reflexive comprehension. This model needs to accomplish five things: (1) It needs to accommodate the target-match asymmetry discussed above. In on-line reading measures, lure-match facilitation is observed in target-mismatch configurations, but not target-match configurations. (2) It needs to account for the variable influence of verb type on lure-match facilitation. Speech verbs produce more lure-match facilitation, but only when the embedded subject is animate. (3) It needs to derive blocking effects. (4) It needs to allow that logophlexive interpretations persist in off-line judgments, even in target-match configurations. (5) It needs to explain why local binding is still generally preferable.

The model proposed here will be based, primarily, on the cue-based retrieval implementations of reflexive reference implemented in Kush (2013) and Parker and Phillips (2017). It augments these models by positing a logophoric operator (OP_{log}) in the left periphery of embedded clauses. This operator refers to the perspective center, and may locally bind the embedded reflexive, giving rise to apparently non-local reference. In what follows, I first present the basic cue-based models

from which this model is derived, and then show how the addition of OP$_{log}$ can derive desiderata (1)-(5), above.

5.2.1 Incorporating logophoricity into comprehension models

In the main, the analysis I propose is situated within the ACT-R framework of sentence comprehension, advanced primarily by Lewis and Vasishth (2005)[3]. It will be useful, then, to remind ourselves briefly of the central components of this framework. The ACT-R framework divides memory into two partitions: procedural memory, and declarative memory. Procedural memory is the component responsible for structure building, and instantiates a left-corner parser operating over a probabilistic context free grammar[4]. In contrast, declarative memory contains all non-procedural information, including objects built by the procedural system. Critically, the workspace for procedural memory is hypothesized to be quite limited, meaning that information not currently being attended must be retrieved from declarative memory if it is to be used for a given procedural operation. So, for example, if the parser is attending to a prepositional phrase, and needs to attach it into the parse it has been building, it must query declarative memory for the relevant syntactic node, re-activate that node to bring it into focal attention, and then apply procedural knowledge to appropriately merge the re-activated node with the currently attended PP. In actuality, then, the driving, explanatory force within the ACT-R framework is its characterization of the retrieval operation which serves to re-activate information for procedural operations. This retrieval mechanism probes memory using a set of "retrieval cues" (roughly, linguistic features) specified by the currently attended input. Constituents in memory which match these cues receive a boost in activation, the size of which is inversely proportional to the number of constituents which match the cue set (the more matches, the more diffuse the associated activation). The probability of retrieval is then a function of activation, such that the likelihood of any one constituent in memory being retrieved is a function of (1) how well that constituent matches the probe, and (2) that constituent's base activation. If a constituent is already fairly active, *and* receives a relatively large activation boost from the retrieval probe, then it is extremely likely to be retrieved. Finally, the framework incorporates a decay function, such that once an item has been moved from procedural to declara-

[3]There is no principled reason for choosing this framework over related theories (McElree, 2000; Van Dyke, 2007; Jäger et al., 2017), but the differences between these implementations (e.g. direct vs. activation-based memory access) should be largely irrelevant for present purposes. The analysis proposed here relies primarily on an elaboration of the syntactic representation. Consequently, it should be reasonable portable among frameworks.

[4]In addition to whatever other procedural operations may be required (e.g. setting the cues for retrieval probes, updating representations, thematic integration, etc).

tive memory, its activation rapidly decays until it is retrieved again. This has the effect of deriving a recency advantage: constituents which were only recently moved to declarative memory will have had less time to decay, and therefore be associated with higher base-activation should they need to be retrieved again. That said, decay is hypothesized to occur quite rapidly, such that the recency advantage is relatively short lived.

With this background in place, we can see how the ACT-R framework has been used to explain antecedent identification in the prior literature. Parker and Phillips (2017), for example, suggest that reflexive pronouns query memory using a retrieval probe specifying morphosyntactic features and structural cues corresponding to Binding Theory constraints. As seen in (105), a reflexive searches memory for antecedents which match its morphosyntactic features[5], and the structural feature LOCAL:1, intended as an implementation of Principle A. In this instance, I adopt the notion of LOCAL advanced in Kush (2013). This feature is dynamically updated to index whether a given DP is (1) a member of the current clause, (2) on the clausal spine. Thus, it collapses the c-command and locality constraints into a single feature acting as a proxy for Principle A of the Standard Binding Theory. An account of this system is given in greater detail in Section 5.3.

(105) **Retrieve:**

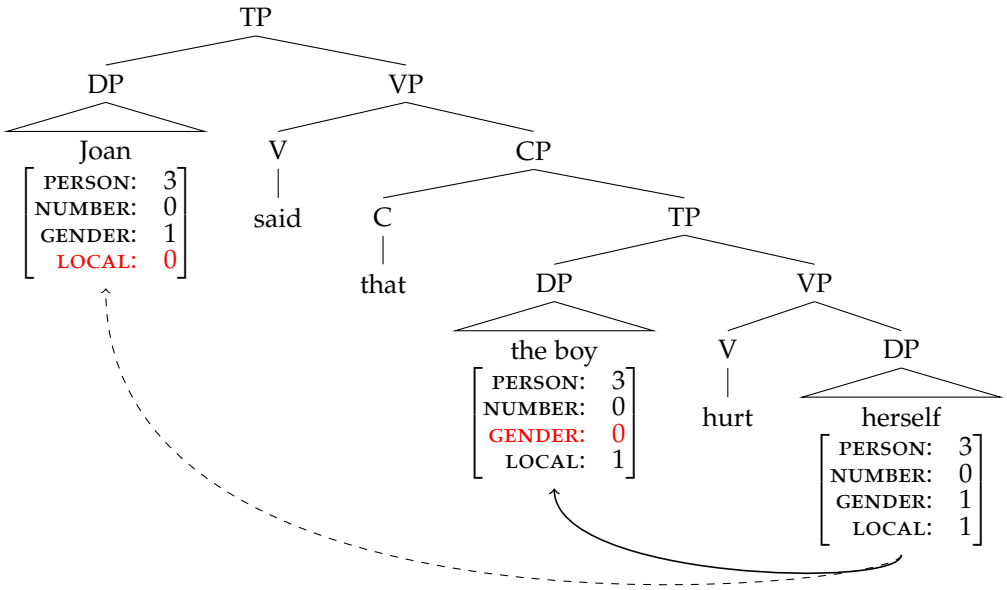
$$\begin{bmatrix} \text{PERSON:} & \alpha \\ \text{NUMBER:} & \beta \\ \text{GENDER:} & \gamma \\ \text{LOCAL:} & 1 \end{bmatrix}$$

Using this probe, Parker and Phillps account for the fact that they only observe lure-match facilitation in the face of gross-feature mismatch by suggesting that the structural feature LOCAL is weighted more highly than match with any one of the probe's morphosyntactic features (i.e., imparts more activation). Thus, when a target antecedent mismatches the probe in only a single morphosyntactic feature, it is still receives relatively more activation because it matches the feature LOCAL:1. However, when a target mismatches in more than one morphosyntactic feature, the amount of associative activation it receives is correspondingly decreased, making lure antecedents relatively more competitive. This process is illustrated graphically in (106a) and (106b). In (106a), neither the embedded, nor the matrix subject is a perfect match for the embedded reflexive. However, the embedded subject only mismatches on the dimension of gender, while the matrix subject fails to satisfy the locality constraint. Given the target's relative recency, and due to the fact that LOCAL imparts more activation than GENDER, the embedded subject has the edge and is more likely to be retrieved than the matrix subject even though it is an imperfect match. In (106a), this

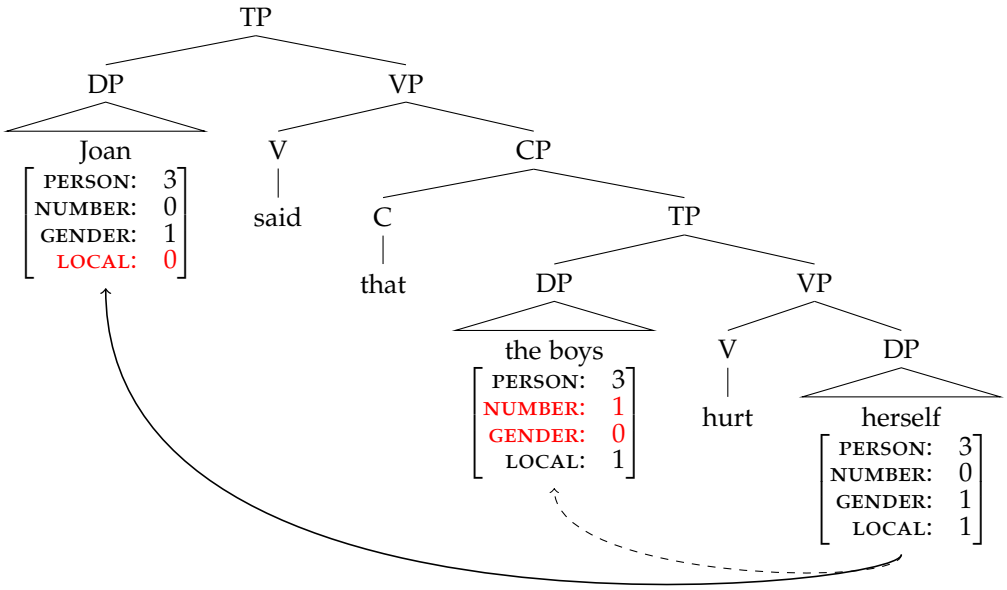
[5]Note that in these representations, NUMBER:1 represents *plural*; GENDER:1 represents *feminine*, and PERSON is numbered according to $1^{st}/2^{nd}/3^{rd}$ person. Alternative specifications may exist, but I do not explore them here.

is indicated with the solid/dashed line distinction. However, in (106b), the embedded subject mismatches the retrieval probe in *two* morphosyntactic features. By hypothesis, this decreases the associative activation of the embedded subject to a degree which allows the matrix subject to be re-accessed, giving rise to Principle A fallibility.

(106) a. **Principle A Adherence** (Parker & Phillips, 2017)



b. **Principle A Fallibility** (Parker & Phillips, 2017)



While compelling, there are several empirical points on which the findings of this dissertation fail to align with Parker and Phillips' treatment of reflexives. First, the findings of Experiment 1 represent a direct contradiction to (at least one) interpretation of Parker and Phillips' findings: gross morphosyntactic mismatch is not always enough to produce lure-match facilitation. Recall that target mismatch in Experiment 1 was achieved with a mismatch in both number, and gender, as it was in Parker and Phillips' studies. However, this experiment only found lure-match facilitation when the embedding verb was a speech verb, not a perception verb. This means that, at the very least, additional factors (e.g. verb thematic role) would need to interact with morphosyntax in Parker and Phillips' system to accommodate these findings.

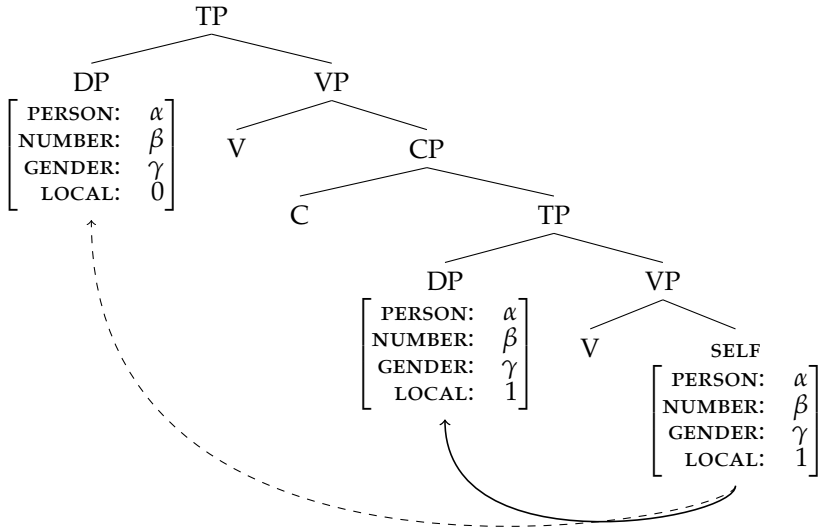
Perhaps more problematic, Experiments 3 and 4 produced two results which conspire to make a cue-weighting story difficult to maintain. First, across all five studies in these experiments there was a consistent main effect of person, indicating that target-mismatch in person is significantly worse ("grosser", to use our previous term) than mismatch in either number, or gender. Under Parker and Phillips' model, this should, descriptively, lead us to expect more sensitivity to lures after an indexical pronoun—grossest feature mismatch should produce the *most* lure-match facilitation. Instead, these studies consistently found *less* lure-match facilitation in the presence of an indexical pronoun, suggesting that the perceived unacceptability of a target/reflexive pair is not directly related to the degree of lure-match facilitation (if anything, it was the reverse).

Admittedly, it might yet be possible to rescue the Parker and Phillips model. One might attempt to modify the cues engaged by reflexives to search for an antecedent such that there is more overlap between a 3sg reflexive and a first person pronoun, than there is between a 3sg reflexive and a 3pl pronoun (for example). This would have the effect that, as far as the retrieval operation is concerned, first person pronouns present *less* of a cue-mismatch than other intervening referents, and thereby predict "person blocking" within Parker and Phillips' system. I cannot rule this out. However, such a solution would then need to explain why a mismatch in person is *perceived* to be so much worse, given that this mismatch actually represents fewer deviations from the retrieval probe. More conceptually, this amendment to the model would fundamentally lose the intuitive generalization that gave rise to it. The "goodness of fit" of a reflexive and its target would not longer be the driving force behind lure-match effects. Instead, a somewhat arbitrary feature specification (which itself mismatches post-retrieval intuitions) would be the critical explanatory factor.

These empirical issues notwithstanding, Parker and Phillips' model provides a useful base from into which we can begin incorporating logophoricity. For ease of exposition, I schematize

their system as shown in (107), where the entry SELF indicates a reflexive pronoun which probes memory for a local antecedent (LOCAL:1) which matches its ϕ features (α, β, and γ).

(107) **Parker and Phillips (2017) (simplified)**



Now, to accomplish the tasks set out at the beginning of this section, I propose an extension of this model building on syntactic theories of logophoricity. Specifically, I follow several syntactic accounts of logophoricity (Charnavel & Zlogar, 2015; Charnavel & Sportiche, 2016; Anand, 2006; Koopman & Sportiche, 1989), in suggesting that logophoric reference is locally mediated by a logophoric operator in the left periphery of embedded clauses[6] (OP$_{log}$), which refers to the "perspective center" of the utterance (Sells, 1987; Anand, 2006; Speas & Tenny, 2003; Huang & Liu, 2001). Thus, *all* instances of reflexive reference are syntactically local, and apparently long-distance interpretations are result of binding via OP$_{log}$. The explanations for the target mismatch asymmetry, variable effect of verb-type, and blocking effects arise from the composition and referential preferences of OP$_{log}$. I address each of these questions in turn.

5.2.2 Accounting for the target-match asymmetry

Before incorporating OP$_{log}$ into the model, more needs to be said about which features are associated with it. Is it endowed with a full complement of ϕ features (perhaps inherited from

[6]However, note that alternative models of licensing logophoric interpretations exist. For example, Pearson (2015) adopts a model proposed by Heim (2001) and von Stechow (2002) (similar in spirit to the model proposed by Culy (1997)) which holds that uninterpretable features on attitude verbs license embedded logophors. As we will see shortly, the availability of OP$_{log}$ as a potential antecedent will be criticla to the model I propose, and so I set these alternatives aside for the present.

its referent), or does it simply carry an index? On the one hand, it might seem simplest to treat OP$_{log}$ like any other potential referent for a reflexive, allowing it to match both a reflexive's ϕ and structural features. This would allow OP$_{log}$ to be highly visible for the purposes of reflexive antecedent identification. However, as seen in (108a), this predicts the wrong state of affairs for preferred reference. If OP$_{log}$ has valued ϕ features, then it is equally as good as any other (ϕ-feature matched) local DP as an antecedent. This should lead us to expect fairly tight competition between local and non-local antecedents (as mediated by OP$_{log}$). Instead, all four eye tracking studies presented here failed to find evidence that non-local antecedents impact the processing of locally-matched reflexives. Likewise, Parker and Phillips (2017) themselves only found evidence of lure-match facilitation when the target was a particularly poor ϕ feature match. Thus, adopting a feature specification like (108a) seems unpromising as a means of capturing the target-mismatch asymmetry.

(108) a. **A local DP and op$_{log}$ with valued ϕ**

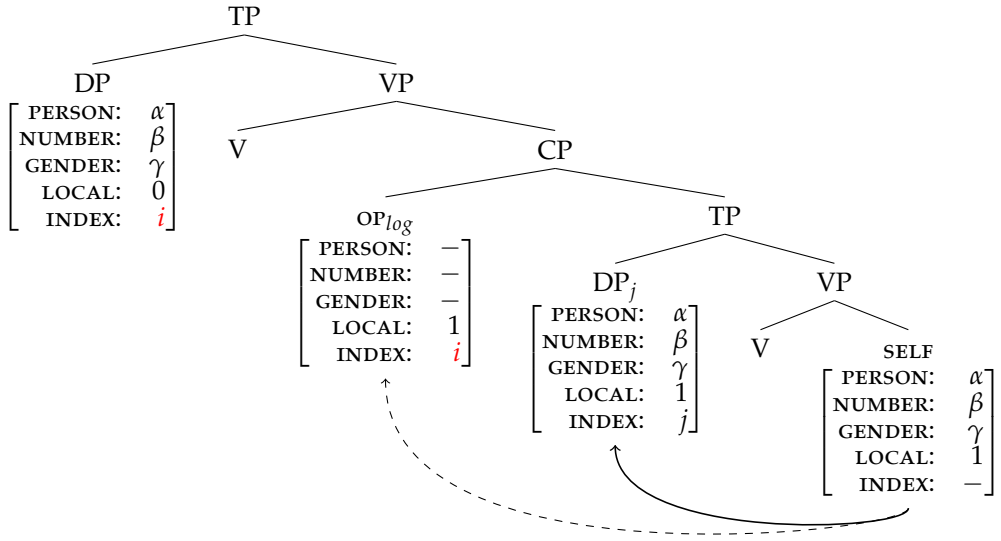
Retrieve:		**DP**		OP$_{log}$	
PERSON:	α	PERSON:	α	PERSON:	α
NUMBER:	β	NUMBER:	β	NUMBER:	β
GENDER:	γ	GENDER:	γ	GENDER:	γ
LOCAL:	1	LOCAL:	1	LOCAL:	1

b. **A local DP and op$_{log}$ with unvalued ϕ**

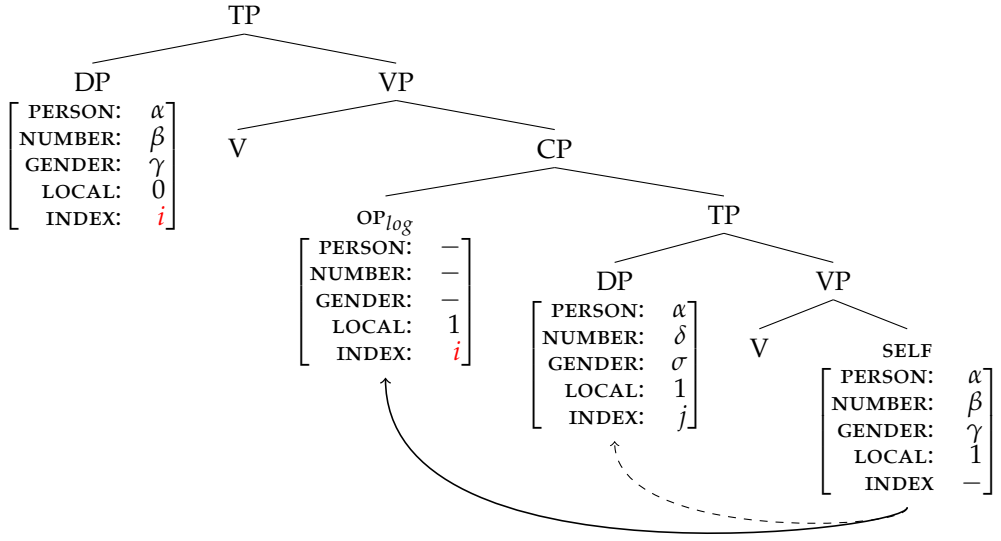
Retrieve:		**DP**		OP$_{log}$	
PERSON:	α	PERSON:	α	PERSON:	$-$
NUMBER:	β	NUMBER:	β	NUMBER:	$-$
GENDER:	γ	GENDER:	γ	GENDER:	$-$
LOCAL:	1	LOCAL:	1	LOCAL:	1

An alternative would be to allow OP$_{log}$ to remain underspecified for ϕ features, as shown in (108b). Given this representation, OP$_{log}$ only matches the reflexive's retrieval probe in a single feature (LOCAL:1), making it a comparatively poor match relative to the local DP (though, importantly, it still does not *mismatch* the retrieval probe). However, when the local DP's ϕ features mismatch those of the probe, the relative activation it receives will be reduced, allowing OP$_{log}$ to emerge as a more competitive antecedent. Thus, by keeping OP$_{log}$'s featural composition fairly sparse, we can explain the differential impact of non-local antecedents on locally matched, and mismatched reflexive comprehension. This contrast is represented graphically in (109) and (110). Note that the representation of antecedents in these trees has been augmented to include an index feature, indicating referents which co-refer. The reflexive itself also bears an index feature, but this feature is unvalued (the reflexive is, after all, searching for an antecedent).

(109) **Logophlexive: target-match**



(110) **Logophlexive: target-mismatch**



The representation of OP$_{log}$ proposed here raises at least two questions deserving closer scrutiny. First, given the results of Experiment 1c, we know that non-local interpretations are possible even in the face of local-match. This suggests that OP$_{log}$ may sometimes be retrieved even when the local antecedent is a perfect ϕ-match for the reflexive. Here it is important to note that, due to the nature of retrieval, it is still technically *possible* for OP$_{log}$ to serve as the antecedent in local-match sentences, simply less *probable*. The relatively high activation of the local DP will render it much more likely to be retrieved, but won't, necessarily, strictly enforce this outcome. Therefore,

we should expect greatly reduced access to OP_{log} in the presence of local-match, but perhaps not a categorical inaccessibility. This seems to be empirically verified. While we found evidence of non-local interpretations in these situations, it remained the dispreferred interpretation ($\sim$30% non-local interpretations in Experiment 1c). The model thus seems to predict the correct outcome.

The second question concerns how non-local mismatch is assessed if OP_{log} itself lacks ϕ features. Comprehenders seem to be sensitive to non-local (mis)match, but it isn't imediately clear how to account for this fact given the absence of syntactic features on OP_{log}. Here I must appeal to whatever mechanisms comprehenders generally employ in evaluating the person/gender/number congruence of a syntactically unbound pronominal and its antecedent. That is, whatever process is used to ensure that the antecedent of *him* in a sentence like *Maeby kissed him* is third person, male, and singular in nature[7] is employed in checking the congruence of OP_{log} and a reflexive[8]

5.2.3 Accounting for verb-type and blocking effects

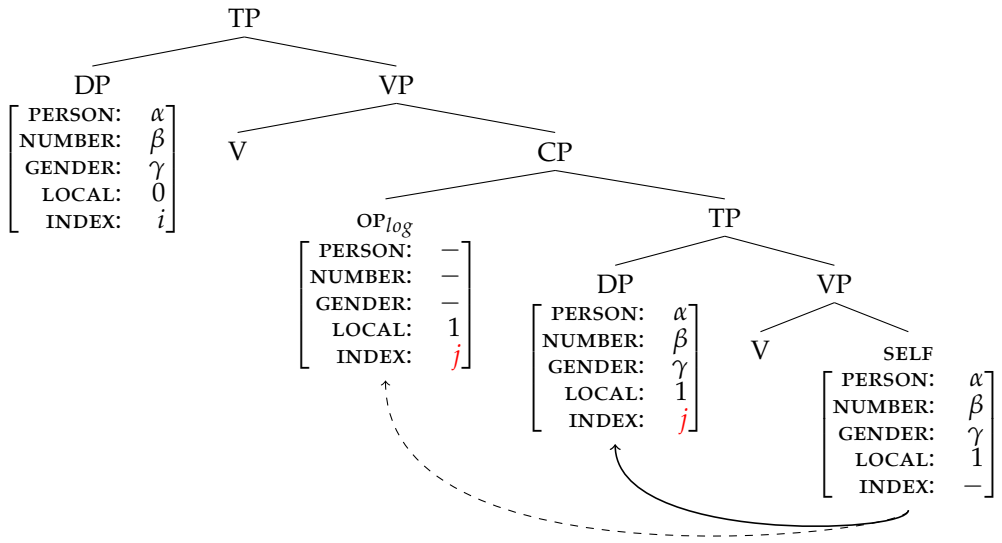
While appropriately specifying the featural composition of OP_{log} nicely derives the target-match asymmetry, it doesn't yet account for the variable effect of verb-type or blocking effects generally. The general form of solution proposed is quite simple: non-local reference is blocked when OP_{log} and the non-local referent do not co-refer (or, more narrowly, when OP_{log} and a local DP co-refer). Schematically, this is the situation shown in (111), where the referents of OP_{log} and the local DP are co-extensive. In this case, either local referent may antecede the reflexive with identical results, and the reflexive cannot refer to the matrix subject.

---

[7] Alas, George-Michael, you are not.

[8] One potentially interesting prediction of this explanation is that it should take longer to evaluate the the feature-match of non-local antecedents than that of local antecedents, as the former must be mediated by the discourse. At present, sufficient data does not exist to evaluate this possibility, and so I leave it as a question for future work.

(111) **No Long Distance Binding**



To achieve the outcome shown in (111), I propose a new model of the mechanism by which OP$_{log}$ finds its referent. The trick lies in deriving the fact that (*i*) speech verb subjects are more likely to co-refer with OP$_{log}$, (*ii*) perception verb subjects are more likely to co-refer with OP$_{log}$ when the embedded subject is inanimate, (*iii*) indexicals are likely to shift the reference of OP$_{log}$. I suggest that all three of these facts derive from the mechanism by which OP$_{log}$ finds its referent. This mechanism has two components, one structural, and one probabilistic. First, I suggest that OP$_{log}$ refers to the highest specified discourse role on Sells (1987)'s hierarchy. That is, OP$_{log}$ must refer to either the SOURCE, SELF, or PIVOT of a clause, and refers as high as possible on this scale. Following Sells, I assume that these discourse roles are hierarchically organized, such that if SOURCE refers sentence internally, then so must SELF and PIVOT. Thus, if SOURCE refers sentence-internally, then OP$_{log}$ refers to SOURCE, and also to the SELF and PIVOT. However, if only the PIVOT role has been sentence-internally specified, then OP$_{log}$ simply refers to the PIVOT. Second, I differ from Sells in suggesting that sentence internal referents are probabilistically mapped onto this scale. For example, attitude holders (the subjects of speech/belief predicates, say) are quite likely to be mapped onto the SOURCE role, and thereby act as the antecedents of OP$_{log}$. Sometimes, however, attitude holders may be mapped to SELF, instead. In this case, if another referent is encountered and mapped to SOURCE (i.e. *above* the attitude holder on the scale), the value of SELF is overwritten with the new value of SOURCE (and likewise with PIVOT), and the reference of OP$_{log}$ is correspondingly shifted.

Using these two components, we can derive each of the effects in Experiments 1-5 by assuming probability distributions over the discourse roles assigned to referents in each case. A hypothetical sketch of this proposal is given in Figure 5.1. Here, I assume that different kinds of referents (e.g. speech-verb subjects, indexicals, and bare animates) are associated with different probability distributions over Sells' discourse role hierarchy. Some antecedents (e.g. speakers) are more likely to control SOURCE, while others are more likely to control PIVOT. However, if a sentence-internal antecedent successfully controls a higher-level role, it obligatorily overwrites the referents of the roles beneath it. The distributions sketched here are derived from the empirical observations seen in Experiments 1-5, as well as the various observations recorded in the syntax literature. As Sells (1987) and Culy (1997) note, *speakers* are far more likely to bind a logophor than *perceivers*. Therefore, it makes sense to center the distribution for *speaker* referents closer to SOURCE, and *perceiver* referents closer to PIVOT. In the middle we have indexical pronouns, which are preferentially anchored to the SELF role, but may probabilistically control either SOURCE or PIVOT. This captures the intuition in (Huang & Liu, 2001) and (Anand, 2006) while allowing for potentially gradient effects of person blocking: indexical pronouns are more likely to be the highest specified discourse role in the hierarchy, but they are not guaranteed to be.

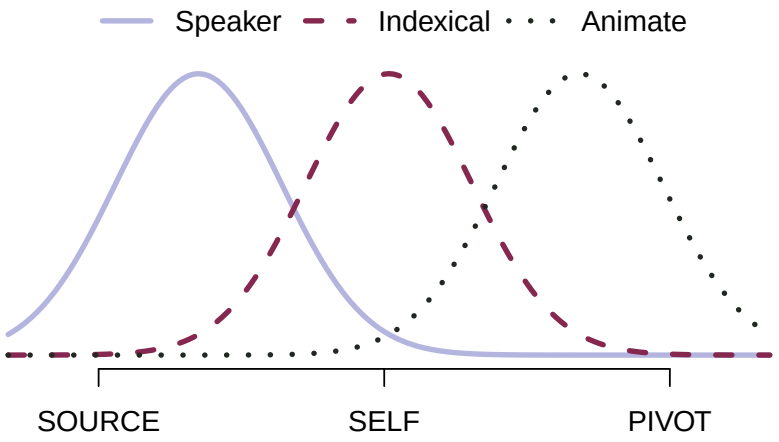


Figure 5.1. Hypothetical probability distributions over antecedents assigned to discourse roles

Assuming these distributions, we can begin to understand the effects of "perspective shifting" seen in Experiments 1-4. In Experiment 1, a speech verb subject is highly likely to control SOURCE, and therefore extremely likely to be co-referent with OP_{log}. In contrast, perception verb subjects are merely animate referents, and therefore much more likely to control the PIVOT. Moreover, upon encountering the embedded subject, control of the pivot is liable to shift to this new animate referent. Therefore, OP_{log} no longer refers to the matrix subject, and long-distance reference is impossible. However, in Experiment 2, this switch in control of PIVOT is obviated: the embedded subject was inanimate, leaving control of PIVOT with the matrix subject. Thus, variability in the effect of verb-type is due to a combination of (1) speech-verb subjects being highly likely to control SOURCE, and (2) perception verb subjects maintaining control of PIVOT in the absence of another consciousness center.

Likewise, we can now extend a similar explanation to person blocking effects. Matrix speakers are only *likely* to control SOURCE. Some proportion of the time they will instead control SELF. Conversely, indexical pronouns are more *likely* to control SELF, but sometimes wind up controlling SOURCE instead. Under this system, person blocking arises when a speech verb subject only controls SELF, and a later indexical binds SOURCE, over-writing the referents of SELF and PIVOT in the process. In this case, OP_{log} will refer to the indexical, thereby preventing long-distance interpretations. Importantly, as described here, the model neatly captures the gradient nature of blocking. As a given referent becomes more likely to control SOURCE, the probability of that referent blocking a long-distance interpretation increases. Thus, we correctly predict that even intervention by animates should, sometimes, give rise to a variety of blocking effects, albeit less frequently.

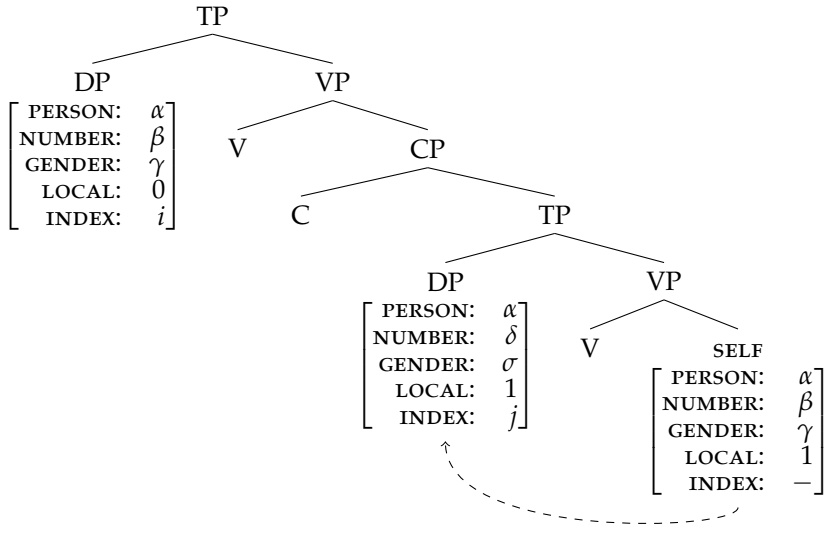
To review: the model I am proposing adopts as a basic premise the ACT-R implementation of antecedent identification articulated in Kush (2013) and Parker and Phillips (2017) (i.m.a.). This model probes memory for LOCAL antecedents which match a reflexive's ϕ features. In addition, I suggest incorporating a ϕ-deficient operator (OP_{log} in the left-periphery of embedded clauses (Charnavel & Sportiche, 2016; Anand, 2006; Koopman & Sportiche, 1989). This operator refers to the highest specified role on Sells (1987) discourse role hierarchy, onto which animate referents are probabilistically mapped as they are encountered. When the matrix subject controls the highest of these discourse roles, OP_{log} may locally bind the reflexive, giving rise to the appearance of long-distance binding. Thus, the model gives a unified explanation of the effects reported in this dissertation: the target-match asymmetry arises from OP_{log}'s ϕ-deficiency, while verb-type, and person blocking effects derive from the probability with which referents are mapped onto Sells'

discourse role hierarchy. Next, I briefly consider two alternative explanations of these findings, and show how they fall short.

5.2.4 Verb-type and person blocking: Alternative analyses

The model proposed above locates verb-type and person blocking effects in the process by which OP_{log} finds an antecedent. However, there may be other ways of modeling these effects. First, one possible alternative explanation for differences between kinds of attitude predicates derives from an Anand (2006)'s suggestion that OP_{log} is only optionally embedded under attitude predicates in English. If so, then we might suppose that some predicates (e.g. speech predicates) are more likely to embed OP_{log} than others (e.g. perception predicates), meaning that perception verbs default to a structure like (112). In this case, even though the local DP is a poor ϕ-match for the reflexive, there is no OP_{log} in the structure to facilitate long-distance reference. Thus, the comprehender is left with the simple percept of ungrammaticality due to target-mismatch.

(112) **No Long Distance Binding (alternative model)**

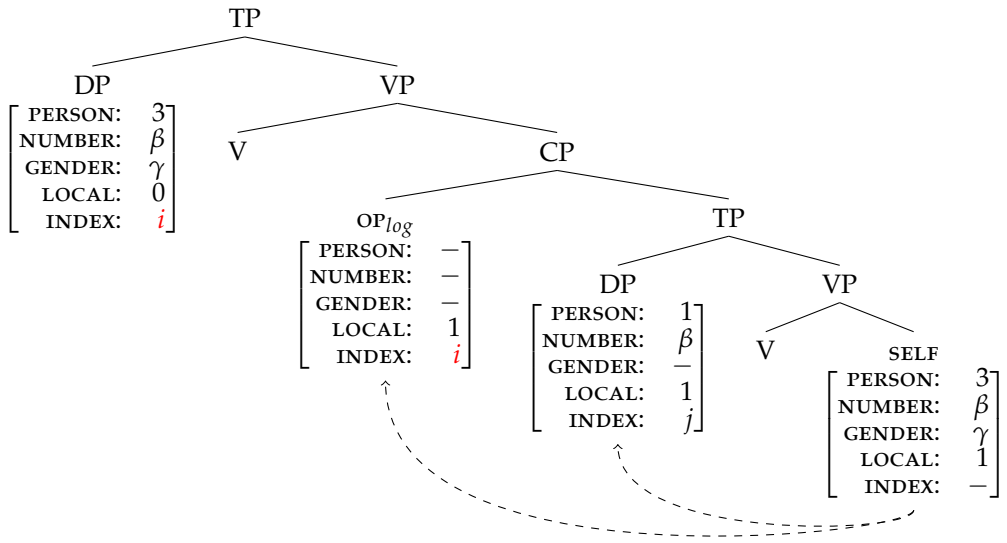


Such a model would struggle to accommodate the contrast between Experiments 1 and 2, however. This contrast critically turned on the nature of the embedded subject: perception verb subjects *do* give rise to lure-match effects when the embedded subject is inanimate. If a structure like (112) is the correct default for perception verb complements, however, then it isn't at all clear why encountering an inanimate embedded subject should trigger retroactive insertion of a logophoric operator. That is, variable insertion of OP_{log} conditioned on verb type doesn't predict (alone) that intervening referents *also* impact the likelihood of lure-match effects. It seems, then, that this alternative model predicts long-distance reference to be variable, but does not capture

the facts for Experiment 2. Moreover, it should not be able to generalize to person blocking effects. In other words, given that we *know* that intervening referents impact the availability of a long-distance antecedent (Experiments 2-4), this model seems to be ruled out from contention.

Perhaps, then, a better tack would be to attempt to account for the impact of intervening reference, before tackling the effect of verb-type. One possible model aimed at capturing such effects was briefly considered in our discussion of Parker and Phillips (2017), above. Recall that, in their model, person blocking (and blocking phenomena in general) can be explained if some embedded subjects share more features with the reflexive than others. One possible instantiation of this intuition would be to suppose that indexical pronouns are underspecified for gender, in which case they only mismatch the retrieval probe in a single feature (PERSON). Under Parker and Phillips' model, then, this should *not* constitute a possible environment for lure-match facilitation (assuming equal weight among ϕ features). However, this solutions suffers from the same fatal flaw we saw previously: it fails to generalize to other instances of blocking. As noted in above, blocking doesn't seem to be a categorical phenomenon, and doesn't seem to be confined to person. While blocking is more effective with indexical pronouns, *any* intervening animate referent seems to block non-local reference to some extent. Consequently, this model seems equally unlikely to account for gradient blocking effects. Moreover, such an account loses the fact that indexicals present a *worse* target-mismatch violation than other local DPs. Finally, it is difficult to see how to generalize it to verb-type effects. In contrast, both verb-type and blocking effects emerge as a natural consequence of the stochastic assignment of referents to Sells' discourse role hierarchy: any intervening referent which overwrites a higher discourse role prevents non-local reference.

(113) **Blocking Effects (alternative model)**



5.2.5 On the longevity of logophlexivity

We come finally to the two remaining desiderata of our model of logophlexivity. The model needs to allow for a durable effect of non-local reference, and explain the general preference for local binding. Happily, both of these conclusions fall out directly from the model we have been examining. First, this model represents a choice among *targets*, rather than an accidental influence of *lures*. When searching for an antecedent, the reflexive isn't considering both embedded and matrix DPs. Instead, it is considering two local antecedents, one of which happens to be co-referent with the matrix subject. Given this, it shouldn't be surprising that logophlexive effects survive in off-line measures. Comprehenders are *literally* considering a grammatical alternative to the more canonical interpretation (as opposed to being hoodwinked into temporarily ignoring an ungrammatical structure). This, then, raises the question of why non-local reference seems to be so difficult, and generally dispreferred to local binding: intuitively, most English speakers seem to feel that local binding is the preferable interpretation.

First, the model already provides one possible analysis of this preference. As noted above, OP_{log} is a relatively poor feature-match for the reflexive (relative to a ϕ-matched local antecedent). Given this, we predict a fairly strong preference for local reference, barring either an absence of a locally-matched DP, or else some additional pressure to adopt OP_{log} as the antecedent. With respect to this latter point, we might appeal to the constraints suggested in (Charnavel & Sportiche, 2016) and Ahn (2015): non-local interpretations may be subject to particular prosodic restrictions

which most direct-object anaphors do not satisfy (without effort). In addition, this model makes available processing-based explanations of the preference for local interpretations of argument reflexives. First, the DP antecedent is relatively more recent than the OP$_{log}$ antecedent, which may lead to a recency preference for the local DP[9] Second, note that local DP (the subject, in these cases) will have been re-activated at the verb for thematic integration, making it highly accessible for subsequent reference. This comparatively high resting activation could render other antecedents relatively inaccessible, leading to an overall dispreference for non-local reference. A similar observation has already been made by King et al. (2012), whose data are presented in greater detail in 5.3.1. A related point comes from converging experimental evidence that adopting local referents is simply easier for reflexives in non-exempt positions (Cunnings & Sturt, 2014; Piñango & Burkhardt, 2005), congruent with the suggestion that activation of the local subject may lead to a processing bias for this interpretation. Finally, it may be that animate embedded subjects are simply too likely to control high positions on Sells' discourse role hierarchy, leading comprehenders to reject non-local antecedents which are not heavily contextually supported as SOURCE referents. Such constraints certainly seem to be in play in the examples in (104), above. In these cases, the long-distance interpretation of reflexives seems to benefit from both prosodic framing, and from fairly strong contextual cues about whose perspective is relevant.

5.2.6 Model Summary

Wrapping up, the model we have converged on consists of the following components: attitude predicates embed OP$_{log}$ in their complements, allowing reflexives to be locally bound by an antecedent which itself refers non-locally. OP$_{log}$ refers to the highest specified level of Sells (1987)'s discourse role hierarchy. Variability in the availability of non-local reference arises from the probabilistic manner in which referents are mapped onto this scale. In addition, an overall preference for local-interpretations results from the relative underspecification of OP$_{log}$, and additional pressures (prosody, base-activation) on non-local interpretations. In the remainder of this chapter, I consider the implications of this model to existing literatures on Binding Theory in sentence comprehension and syntactic theory.

[9]Though note that simulations would be needed to confirm this, especially as decay is posited to be quite rapid.

5.3 Logophlexivity in reflexive comprehension

The model proposed above is based on a cue-based retrieval implementation of antecedent identification (Parker & Phillips, 2017), but incorporates novel extensions to the syntactic search space of reflexives. In building this model, I have already shown how the proposed extensions explain the results of Experiments 1-5, as well as the findings of Experiments 1 and 3 from Parker and Phillips (2017). How well, then, does this model generalize to other findings in the sentence comprehension literature? The present section addresses this question before turning to a consideration of alternative characterizations of Principle A fallibility, with a focus on comparisons to the literature on "grammatical illusions". From there, the section concludes by probing the role played by the cue-based retrieval mechanism in implementing logophlexivity.

Returning to the discussion of reflexives in sentence comprehension from Chapter 1, recall that, until quite recently, the majority of findings indicated that reflexive pronouns strongly applied Principle A at the earliest stages of antecedent identification. In fact, with the exception of Parker and Phillips (2017), previous studies have almost exclusively found little to no evidence of sensitivity to lures in reflexive comprehension. Fortunately, this pattern of behavior is entirely consistent with the logophlexives model—previous studies which failed to find lure-match effects either (*i*) used lures which were poor antecedents for OP$_{log}$, and/or (*ii*) configurations in which OP$_{log}$ could not effective out-compete local antecedents.

For example, in the case of (Sturt, 2003b), the lure was only rarely the subject of a speech predicate, and the embedded subject was animate. Given this, the logophlexives hypothesis explains the lack of lure-match effects in his studies as the product of the relatively low probability of lures controlling SOURCE, and the relatively high probability of targets controlling at least the PIVOT. Likewise, studies like Xiang et al. (2009) and Dillon et al. (2013) embedded lures inside relative clauses, making them unlikely to control any discourse role other than, perhaps, PIVOT. Finally, the cross-modal semantic priming results obtained by Nicol and Swinney (1989) can be explained by appealing to the ϕ deficiency of OP$_{log}$. In the absence of morphological pressures pushing comprehenders to consider OP$_{log}$, non-local referents will not be reaccessed, in which case they will not produce semantic priming effects[10]. Setting aside non-argument reflexives for the present (these cases will be considered below, in Section 5.4), then, the logophlexives model

[10] An alternative analysis here might suggest that antecedent re-access mediated by OP$_{log}$ is distinct from directly accessing an antecedent. If so, it may be that the lemma-level information associated with the lexical antecedent *isn't* reaccessed when interpreting OP$_{log}$. If so, we would not expect to observe semantic priming from non-local antecedents even when comprehenders adopt a non-local interpretation. This seems like a somewhat baroque explanation, however, given that the logophlexives model can already accommodate Nicol and Swinney (1989)'s findings, and so I do not pursue it further.

seems to handle the distribution of findings in the previous literature quite well. As expected, sensitivity to non-local referents arises when those referents are likely to be co-referent with OP_{log}, and when OP_{log} itself is more viable than other local referents. Next we consider cases which the logophlexives model does not so neatly explain.

5.3.1 What logophlexivity doesn't explain

Despite the wide scope of effects explained by the logophlexives hypothesis, there remain a few points which continue to prove troubling. The first of these is Experiment 2 in Parker and Phillips (2017), which investigated sentences like (114)[11]. Unlike in their other experiments, the target in this study was the subject of the main clause, and the reflexive was the main-clause direct object. The reflexive and target always mismatched, as the target was always singular, and inanimate (e.g. *the soothing tea*), while the reflexive was alway animate, and plural (i.e. *themselves*). The lure, then, was the subject of an object relative clause modifying the target. Lures were manipulated so that they either matched, or mismatched the reflexive in number (*students*/*student*). The task, as before, was eye-tracking while reading.

(114) $*$ The soothing tea that the nervous $\left\{ \begin{matrix} \text{student} \\ \text{students} \end{matrix} \right\}$ drank calmed themselves down...

Mirroring their other experiments, Parker and Phillips again report substantial lure-match facilitation in these sentences: reflexives were read more quickly when they matched the lure. On its face, this finding constitutes a problem for the logophlexives hypothesis. If sensitivity to lures only arises in logophoric contexts, then we would be forced to argue that the subjects of object relative clauses can act as logophoric antecedents. Given that logophoric antecedents tend to be the subjects of attitude predicates (Sells, 1987; Culy, 1997; Pearson, 2015; Speas & Tenny, 2003), this seems to be *a priori* unlikely. However, this configuration does bear remarkable similarity to the "sub-command" configuration described for *ziji* in Mandarin. In this configuration, an animate subject contained inside an *inanimate* subject, may bind a non-c-commanded instance of *ziji* (Huang & Liu, 2001; Tang, 1989), as seen in (115).

(115) a. Zhangsan tou dongxi de shishi bei ziji de laoban faxian le.
Zhangsan steal things DE fact BEI SELF DE boss discoverb PERF.
The fact that Zhangsan$_i$ stole things was discovered by his$_i$ boss.

b. Zhangsan$_i$ de biaoqing gaosu wo$_j$ ziji$_{i/*j}$ shi wugude.
Zhangsan$_i$ DE expression tell me$_i$ SELF$_{i/*j}$ is innocent

[11] This experiment included no grammatical controls for reflexive conditions.

Zhangsan$_i$'s expression tells me$_j$ that he$_{i*j}$ is innocent.

c. $*$ Zhangsan$_i$ de shibai biaoshi tamen dui ziji$_i$ mei xinxin.
Zhangsan$_i$ DE failure indicate they to self$_i$ no confidence
Zhangsan$_i$'s failure indicates that they have no confidence in him$_i$.

That said, this is one surprising use of *ziji* to which Huang and Liu (2001) do *not* attribute a logophoric source. Unlike other cases, they note that sub-commanded ziji does not exhibit blocking effects, as seen in (115b), and sub-commanding antecedents cannot bind a non-local *ziji*, as seen in (115c). Thus, sub-command configurations fail two of the diagnostics used to identify logophoric *ziji*. Based on this, Huang and Liu suggest that sub-commanded *ziji* is an instance of local binding, albeit one with a slightly different notion of c-command.

Pulling this back to the findings of Parker and Phillips' Experiment 2, we're left with the conclusion that these findings do not reflect a logophoric use of reflexives, but may reflect the same grammatical principles which allow for sub-command in Mandarin. To assess this, one would want to test whether the principles which govern sub-command hold true in this same configuration in English. Does it require an inanimate subject? Is it subject to blocking effects or locality restrictions? I leave these questions for future work. Regardless, Parker and Phillips' findings accord well with cross-linguistic patterns of behavior, and so deserve closer scrutiny even as that explanation may lie outside the scope of the logophlexives hypothesis.

The second point of contention comes from data reported in King et al. (2012). In yet another eye-tacking while reading study, these authors investigated sentences like (116). In this study, the authors investigated the possibility that the recency of subject re-activation for verb-adjacent reflexives was responsible for the strong effects of Principle A observed in previous work. Recall that the likelihood of a given element being retrieved is a function of (*i*) the degree to which that element matches the retrieval cues, and (*ii*) the base-activation of that element. Critically, when an element is retrieved, it receives an activation boost, making it easier for subsequent retrieval. Reasoning from this model, King et al. suggested that verb-adjacent reflexives will always have a highly active subject available as an antecedent—this subject having just been retrieved for thematic integration with the verb. Thus, the relative infallibility of Principle A may be no more than the product of the recent re-activation of subjects for verb-adjacent reflexives. In their study, they manipulated the position of the reflexive so that it was either verb adjacent, as in (116a), or verb non-adjacent as in (116b). Following previous studies in the literature, they also manipulated whether the reflexive matched or mismatched a target antecedent (the main-clause subject), or a

lure referent (the object of a subject relative clause modifying the target). In first pass, and go-past reading times at the critical reflexive they found a significant lure-match effect for target-mismatch reflexives, but only when the reflexive was not adjacent to the verb. When the verb and reflexive were adjacent, there was no impact of lure-match.

(116) a. The bricklayer who employed $\left\{\begin{matrix}\text{Gregory}\\\text{Helen}\end{matrix}\right\}$ shipped $\left\{\begin{matrix}\text{himself}\\\text{herself}\end{matrix}\right\}$ sacks of mortar...

b. The bricklayer who employed $\left\{\begin{matrix}\text{Gregory}\\\text{Helen}\end{matrix}\right\}$ shipped sacks of mortar to $\left\{\begin{matrix}\text{himself}\\\text{herself}\end{matrix}\right\}$...

Again, these findings seem to be inconsistent with the predictions of the logophlexives hypothesis. Under this model, it isn't immediately clear why distance from the verb should impact sensitivity to lures. Moreover, lures in King et al.'s study were in an even *worse* position than in Parker and Phillips' Experiment 2 to act as logophoric antecedents. As the objects of SRCs, it isn't clear that these lures should be able to antecede a logophoric use of the reflexives. This forces us to one of two conclusions. Either the data of King et al. (2012) represents an as-yet-to-be-understood instance of logophoricity, or it really is an example of a grammatical illusion, as discussed above. If the former, then our model of what kinds of referents are likely to be assigned perspective needs substantial modification. One point in favor of this position is that verb non-adjacent reflexives, as in (116b), may be assigned primary stress, a factor (Charnavel & Sportiche, 2016) and Ahn (2015) associate with non-canonical interpretations. Since King et al.'s items did not involve an attitude holder, this may have allowed the SRC object to overwrite the PIVOT role, and act as the referent for OP_{log}. However, this would require that OP_{log} independently exists in the left periphery of root clauses, a possibility not discussed so far in this work. Even so, this may not be an entirely indefensible position, given proposals like Speas and Tenny (2003), which posit a logophoric operator in the root of all sentences. Taking an entirely different tack, King et al.'s findings could represent a true grammatical illusion, in which case we are forced to conclude that not all instance of Principle A fallibility are grammatically guided. At present, I cannot arbitrate between these two alternatives, and so leave this investigation for future work. Happily, however, King et al.'s results suggest a possible alternative interpretation of the argument/non-argument distinction posited by predicate-based models of binding. Perhaps these models are attempting to capture the intuition that a verb non-adjacent reflexive is more likely to refer to an antecedent which is not an argument of that verb, a hypothesis explored in greater detail in Section 5.4.

In sum, there are two examples of data which do not comport well with the logophlexives hypothesis. In both cases, lure-match facilitation was observed even though lures were not in positions canonically associated with logophoric refence. For Parker and Phillips (2017), at least,

we can appeal to the grammatical constraints which apply to sub-command configurations in Mandarin, although further work is needed to assess this explanation. King et al. (2012) present a more pressing challenge, as their lures were quite bad logophoric antecedents, and their primary effect (the effect of verb non-adjacency) unpredicted under the logophlexives hypothesis. These findings may very well indicate grammatical fallibility on top of the logophlexives effect. For now, I leave this investigation for future work, and turn to a possible alternative characterization of logophlexivity.

5.3.2 Logophlexivity isn't a grammatical illusion

One frequent point of comparison for "Principle A fallibility" is the well-known class of "grammatical illusions"; cases in which structurally illicit elements appear to (temporarily) license otherwise ungrammatical morphology. Perhaps the two most well-studied of these phenomena are agreement attraction and illusory NPI licensing. In agreement attraction, subject-verb disagreement is perceptually rescued when some non-subject noun phrase matches the verb's ϕ-features. An example of this is given in (117), where comprehenders frequently fail to notice that the plural verb *were* is technically ungrammatical (it mismatches its subject, *the key*) because it matches the non-subject noun *cabinets* (Eberhard et al., 2005; Wagers et al., 2009; Parker & Phillips, 2016; Dillon et al., 2013, i.m.a.). Similarly, negative polarity items (NPIs) like *ever* must be in the scope of negation (roughly, c-commanded) to be licensed[12]. Thus, the NPI in sentences like (118a) is unlicensed in the absence of negation. However, as with agreement attraction, comprehenders have been shown to mis-perceive sentences like (118b) as grammatical, due to the presence of structurally inappropriate negation in the preceding context (Vasishth, Brüssow, Lewis, & Drenhaus, 2008; Xiang et al., 2009; Xiang, Grove, & Giannakidou, 2013; Parker & Phillips, 2016, i.a.).

(117) $*$ The key to **the cabinets** definitely **were** rusty with disuse.

(118) a. $*$ The bill that the senators voted for will ever be signed into law by the president.

b. $*$ The bill that **no** senators voted for will **ever** be signed into law by the president.

On the surface, the parallel between these phenomena and the logphlexive data discussed here is clear. In fact, we can even use the same terminology to discuss them. For verb agreement, the TARGET is simply the subject, while other, non-subject noun phrases constitute LURES. Likewise, for NPI licensing, a c-commanding instance of negation is the intended TARGET, and non-

[12]This is a gross over-simplification of NPI licensing, but it will serve the purpose of this discussion.

c-commanding negation would be a LURE. Thus, agreement attraction and illusory NPI licensing seem to present much the same problem as logophlexives: in cases of target-mismatch, comprehenders are willing to consider lure referents to rescue an otherwise ungrammatical sentence. Put this way, logophlexives join the class of "grammatical illusions", representing the misepreception of grammaticality in the face of surface-satisfaction of an element's morphological licensing needs.

Before continuing, I would be remiss in failing to point out that this intuition is common, and represented repeatedly in the literature. Xiang et al. (2009), for example, conducted a direct comparison of reflexive comprehension and NPI licensing in an ERP study. They found considerable evidence that lures decreased the size of N400 for otherwise unlicensed NPIs (relative to completely unlicensed NPIs). However, no such modulation of either the N400, or the P600 was observed for target-mismatched reflexives. Similarly, Dillon et al. (2013) directly compared agreement and reflexives in a series of eye-tracking while reading studies. Replicating both respective literatures, they found significant lure-match facilitation for agreement (i.e. agreement attraction), but only target-mismatch effects for reflexives, with no particular impact of lure. In fact, along with Sturt (2003b), it was partially as a consequence of studies like these that researchers came to believe that Principle A was a remarkably robust constraint on the antecedent identification process (Sturt, 2003b; Dillon et al., 2013; Xiang et al., 2009; Cunnings & Sturt, 2014). Given that we have seen extensive evidence in this dissertation (and in more recent studies (Parker & Phillips, 2017)) that this is not the case, it seems worth re-evaluating whether Principle A fallibility should be treated on par with other grammatical illusions.

In fact, this is precisely what Parker and Phillips (2017) were suggesting with their model. Grammatical illusions like agreement attraction and illusory NPI licensing have frequently been treated as the product of errors in memory retrieval. The parser encounters morphology in need of licensing, and initiates a retrieval to find the target licensor. In the absence of a licit target, this may result in lure licensors accidentally being retrieved, spuriously satisfying the morphological dependency. Parker and Phillips extended this model to reflexives by suggesting that structural cues to antecedent identity are weighted more highly than morphosyntactic cues. However, despite the surface similarity between logophlexivity and grammatical illusions, I would like to argue that the two should not be collapsed into a homogeneous phenomenon. This argument consists of two pieces. First, the character of the explanation assigned to grammatical illusions within a cue-based retrieval framework is markedly different from the character of the explanation provided by the logophlexives model. Second, logophlexive behavior is supported by cross-linguistically attested patterns of behavior, while neither agreement attraction, nor illusory NPI licensing is. In brief,

both of these arguments articulate the view that grammatical illusions represent true failures to apply grammatical knowledge, while logophlexivity is the product of alternative grammatical constraints.

5.3.3 On the contribution of cue-based retrieval

In proposing their model of sentence processing, Lewis and Vasishth (2005) identified three key components: (*i*) the architectural assumptions of ACT-R; (*ii*) a left-corner parsing algorithm; (*iii*) the representations of theoretical syntax. In explaining grammatical illusions, most models have appealed to (*i*). That is, they rely on the architectural principles of ACT-R (in particular, a stochastic retrieval mechanism) to explain deviations from expected behavior.

So, for example, agreement attraction is known to be susceptible to several grammatical influences: it is typically associated only with "ungrammatical" sentences (a "target-match", or "grammaticality" asymmetry), and is more pronounced with plural agreement/lures (a markedness asymmetry). In addition, lures embedded in PP post-modifiers tend to exert a greater influence than lures embedded in relative clauses (Hammerly & Dillon, 2017), and matrix subjects seem to exert little to no influence on embedded agreement (Sturt & Kwon, 2017). In other words, it would be mistaken to say that agreement attraction is not influenced by grammatical factors—it clearly is. Nonetheless, the primary explanatory force assigned to agreement attraction has primarily relied on architectural facts about retrieval. In particular, agreement attraction is hypothesized to arise from the probabilistic nature of cue-based retrieval (Wagers et al., 2009; Dillon et al., 2013; Parker & Phillips, 2017, i.m.a.). That is, in agreement attraction, lures are never retrieved because they perfectly match the retrieval, probe. Instead, they are retrieved because they match some subset of the probe's features *despite* the fact that they mismatch others. In other words, lures represent imperfect, probablistic controllers for agreement.

The situation with logophlexives, at least in the model proposed above, is decidedly different. While the nature of the retrieval mechanism plays a role in explaining the target-match asymmetry, it does relatively little else. Instead, the explanation for verb-type and blocking effects derives entirely from the nature of OP_{log}, and the manner in which referents are probabilistically mapped to Sells (1987)'s discourse role hierarchy. Thus, the primary explanatory force for logophlexivity derives from OP_{log} and its affiliated properties, not from properties of the ACTR-R architecture. That is, the hypothesis explored here eschews a retrieval-based explanation in favor of an expansion of the syntactic representation; positing a logophoric operator as the primary explanatory device.

Notably, while this operator is a *dispreferred* antecedent for a reflexive, it never actually *mismatches* the embedded reflexive, and so constitutes a grammatically possible alternative.

Grammatical illusions and logophlexivity thus find explanations in different aspects of Lewis and Vasishth (2005)'s model, the former deriving from architectural principles of ACT-R, while the latter recruits insights from syntactic theory. Conceptually, this corresponds to treating grammatical illusions like processing errors, while logophlexivity represents grammatical, albeit disadvantaged, behavior. Framework internally, then, we have good reason to differentiate between grammatical illusions and logophlexivity.

5.3.4 On the grammatical nature of logophlexivity

While the two phenomena receive different explanations within the ACT-R framework, this theory-internal distinction should correspond to empirical differences between the phenomena. Perhaps agreement attraction should be treated more like logophlexivity (or vice versa) and our model is making an inappropriate distinction between the two. There are three lines of argumentation against this worry. I list these arguments in brief before engaging with each in more detail. First, as we'll see shortly, the fact that long distance reference is mediated by OP$_{log}$ (as opposed to directly accessing non-local referents) renders logophlexivity and grammatical illusions importantly different. Thus, to the extent that we need to OP$_{log}$, we have evidence that the two phenomena are distinct. Second, the two phenomena actually produce different consequences for sentence interpretation, indicating that the two should be assigned distinct explanations. Finally, patterns of cross-linguistic grammatical behavior support logophlexivity, while few, if any languages grammaticize the behavior observed with grammatical illusions. Consequently, elaborating the grammar to explain logophlexivity seems reasonable, while grammatical illusions appear to be better modeled as retrieval errors.

As discussed above, the critical difference between a phenomenon like agreement attraction and my model of logophlexivity within the ACT-R framework is that agreement lures actually mismatch the retrieval probe on some dimension (e.g. syntactic position), while OP$_{log}$ does not. Thus, agreement lures are *not* grammatical controllers for agreement, while OP$_{log}$ technically *is* a grammatical antecedent for a reflexive. One alternative would be to attempt to unify these phenomena by making logophlexivity more similar to agreement attraction. A possible instantiation of this would be to do away with OP$_{log}$ entirely, and allow that non-local referents may be directly accessed by embedded reflexives. In this case, the non-local referent would be identical to an agreement attraction lure, inasmuch as both represent mismatches with the retrieval probe. How-

ever, as we saw in Section 5.2, there are good reasons not to adopt such a model: abandoning OP_{log} makes it difficult to capture verb-type and blocking effects simultaneously. Barring an alternative model of logophlexivity, then, it seems difficult to reduce it to simply another case of grammatical illusion—at least, to the extent that illusions are operationally defined on the basis of retrieval probe mismatch.

More concretely, grammatical illusions and logophlexivity do not seem to have the same effect on sentence interpretation. At present, it appears that agreement attraction occurs without interpretive consequence. That is, comprehenders don't appear to interpret sentences like (119a) to mean that *the widows were worried*, even though they use the lure's plural feature to check the verb's agreement (Schlueter et al., 2017). Such findings suggest a dissociation between the agreement checking process (and its susceptibility to error) and thematic integration. In contrast, as we saw in Experiment 1c, comprehenders *do* assign non-local interpretations to reflexive pronouns in sentences like (119b): as much as 30% of the time, comprehenders interpret *herself* as referring to *the nurse*. This is notable for two reasons: first, it indicates that non-local reference is not contingent on gross target-mismatch (c.f. (Parker & Phillips, 2017)); and second it suggests that when we observe lure-match facilitation effects for reflexives, comprehenders aren't simply checking that the reflexive's morphosyntactic features are licensed. Instead, they are actively entertaining non-local referents as antecedents. Thus, we have an actual empirical difference in the effect of attraction and logophlexivity, suggesting that they represent distinct classes of behavior.

(119) a. The nurse of **the elderly widows** apparently **were worried** about the risk of infection.

b. **The nurse** said that the elderly widow worried **herself** because of the risk of infection.

Finally, logophlexivity, but not grammatical illusions, is supported by cross-linguistic grammatical behavior. Logophlexivity seems to be a valid, grammatical form of expression in several languages, making it reasonable to suggest a grammatical source of the behavior in English. In contrast, at least the canonical cases of agreement attraction do not seem to be attested grammatical structures across the world's languages. PP post-modifiers do not, as a general rule, control agreement[13]. Similarly, NPIs do not typically seem to be licensed by negation in a non-c-

[13]Note that this may not be true of other structures which give rise to unexpected agreement. For example, Dillon, Staub, Levy, and Clifton Jr (2017) note that comprehenders frequently prefer agreement with a fronted WH object, as in *Which flowers are the gardener planting?*. There may be languages which actually grammaticize this behavior. Under the thesis being explored here, then, this should likewise not be collapsed with "agreement attraction", and instead search for an explanation in terms of an elaborated linguistic representation.

commanding position. Given this, it seems reasonable to attribute logophlexivity to a grammatical source, and grammatical illusions to a processing-based explanation, as suggested above.

In sum, the model of logophlexivity proposed in this dissertation gives a very different characterization of the behavior than the one usually assigned to agreement attraction. Theory internally, then, the two represent distinct phenomena, with unique explanations. Importantly, this theory internal distinction seems to be externally motivated. To collapse grammatical illusions and logophlexivity, we would need to abandon OP$_{log}$, a move which would then lose us the ability to account for verb-type and blocking effect in reflexive comprehension. Moreover, attraction and logophlexivity differentially impact sentence interpretation, indicating a divergent source for the two phenomena. Finally, and perhaps most compellingly, logophlexivity, but not attraction, is a cross-linguistically attested pattern of grammatical behavior, suggesting a grammatical source for the former, but not the latter.

5.3.5 Outstanding issues for cue-based implementations

We turn now to an issue which plagues not only this, but all current retrieval-based models of reflexive antecedent identification: the question of how, exactly, to encode Binding Theoretic constraints as cues to be used in retrieving an antecedent. Up till now, I have largely side-stepped this issue for expositional reasons, relying entirely on Kush (2013)'s LOCAL feature without properly defining it. However, the specification and maintenence of this feature is itself a non-trivially difficult issue which deserves closer attention. This section thus explores exactly why Binding Theory constraints are so problematic for cue-based implementations, and the current state of the field's attempts to address the issue.

One widely acknowledged weakness of cue-based approaches to encoding linguistic knowledge is the difficulty they encounter when confronted with inherently relational constraints. Retrieval cues (the means by which linguistic constraints are, necessarily, encoded) are essentially one-place predicates, tagging a constituent with a particular property. For example, morphosyntactic matching constraints (e.g. agreement) are fairly straightforwardly encoded in a feature-matching search system. Constituents in memory either bear, or lack, the appropriate morphosyntactic features, and are consequently either re-activated or not by the retrieval probe. Similarly, particular structural positions can be straightforwardly targeted: in looking for a subject, a verb my probe memory for things that are [+SpecTP], and inasmuch as a constituent either is, or is not, a specifier of TP, it will be reactivated.

Unfortunately, binding constraints rely on two notions which are not so easily encoded with simple features: locality, and c-command. Consider first locality. In establishing whether a constituent x is local relative to y, one needs to know both the relevant definition of locality (e.g. the local clause), and the relative positions of both x and y. When presented with a fully-formed syntactic tree, it is straightforward to use this information to mark x with the relevant locality value: x is "local relative to$y$" iff x and y are in the same clause. However, when attempting to mark constituents with these features in real time, the centre cannot hold[14]. First, consider that in order to mark x as "local relative to y", we first need to know the position of$y$. However, when x is first encoded (assuming x precedes y), it *cannot* be assigned this feature, as the relevant information is not yet available. Second, updating x with the appropriate value of "local" is non-trivially difficult. The most obvious answer would be to attempt to retrieve x after encountering y so that x's representation may be updated. However the very feature we would like to update ("local relative to y") is exactly the feature we would like to use to retrieve x, and the very feature that x critically lacks. Much the same problem exists for attempting to implement c-command in terms of one-place features. To know whether x c-commands y, one must first know the position of y. Therefore, x cannot have been encoded as c-commanding y prior to encountering y, and [+c-command] cannot be feature used to search for x.

In short, the "structural cues" used to search for a reflexive antecedent are not obviously compatible with the feature-valuation system assumed in cue-based retrieval frameworks. This seems problematic, as the claim that linguistic constraints may be implemented as retrieval cues is dependent on an understanding of how those cues are assigned and valued. A cue-based model of a linguistic dependency is only explanatory inasmuch as the cues employed are justified, and well-understood. Thus, positing cues like [+c-command] or [+local] to explain reflexive antecedent search rather misses the point, as these features *couldn't* be implemented as such. How, then, can we reconcile the insights of cue-based models (a stochastic, associative memory access architecture) with the limitations imposed by linguistic constraints?

One possible solution to this problem has been advanced in Kush (2013), who proposed an algorithmic mechanism for valuing the feature LOCAL (120)[15]. In effect, this mechanism collapses

[14]Which is to say, things fall apart.

[15]This mechanism has a co-argument based variant which assigns the value 1 to a constituent x if x is a coargument of the current clause's verb, and 0/NULL otherwise. Kush (2013) actually prefers this model, on the grounds that the spine-mate model struggles with reflexives in the complements of thematic PPs (e.g. *John talked to himself*). However, a definition of LOCAL relying on co-argumenthood would itself have to distinguish between thematic and non-thematic PPs to capture the facts for Dutch (and exempt anaphors in English). Given this, and the arguments presented in Section 5.4 against an argument/non-argument distinction for reflexives, I set aside this alternative proposal here.

the locality and c-command components of Standard Binding Theory, marking all and only those constitutions which are spine-members of the current clause LOCAL:1. Critically, Kush also proposes a local-update procedure which re-writes the value of LOCAL for all constituents not in the current clause. Thus, in multi-clausal sentences, the matrix subject will not be LOCAL:1 by the time an embedded predicate is encountered. This achieves roughly the desired outcome for reflexive-cued retrievals: a reflexive pronoun will search for clause-mate constituents on the clausal spine (i.e. c-commanding). Moreover, this algorithm provides a clever means of implementing Principle B: pronouns simply search for constituents which are LOCAL:0.

(120) $\textsc{local}(x, \text{current.Clause}) = \left\{ \begin{array}{cc} 1 & \text{if x is a DP on the spine of the current clause} \\ 0 & \text{if x is any other DP} \\ - & \text{if x is not a DP} \end{array} \right\}$

However, this implementation struggles with a few key binding facts, summarized in (121). Broadly speaking, these are the same problems we tackled in Section 1.1 for the Standard Binding Theory: binding domains don't always correspond to the clause. In particular, recall that the referential distribution of reflexives and pronouns overlap when they are not themselves directly on the spine of the clause, as seen in (121). Kush's solution here is to relativize the retrieval probe for pronouns to the LOCAL value of the pronoun itself. If the pronoun is on the clausal spine, (i.e. LOCAL:1), it will engage a retrieval for LOCAL:0 to identify an antecedent consistent with Principle B. If, however, the pronoun is *not* on the clausal spine (i.e. LOCAL:0), then the retrieval probe will be underspecified for LOCAL (LOCAL: $-$), allowing morphosyntax to guide antecedent identification.

(121) a. Lucille$_i$ doesn't like that picture of herself$_i$/her$_i$ in the local news.

b. [George-Michael and Maeby]$_i$ like each other$_i$'s/their$_i$ parents better.

c. The Bluths$_i$ claimed that those stories about themselves$_i$/them$_i$ in the paper were utterly false.

This amendment should produce the correct results for non-argument reflexives and pronouns, but the case in (122) may prove more troublesome. In ECM constructions like this, the embedded subject can be bound by the matrix subject (as in 122a), or, conversely, bind the embedded object (as in 122b). Thus, the critical reflexive in (122a) needs to be LOCAL:1 relative to the matrix clause, even though another DP in the same position needs to be LOCAL:1 relative to the embedded clause. This does not seem to fall out of the LOCAL valuation algorithm in (120), meaning that some amendment to the clause-boundary identification algorithm is needed. One possibility here would be to allow the ambiguity of the role of the embedded subject to play into the valuation of LOCAL. Perhaps ECM objects are effectively "double counted" as LOCAL:1 because they are

treated as the direct object of the matrix verb when first encountered, and only retroactively re-analyzed as the embedded subject. However, this should predict that if the role of the ECM object is *unambiguous* (as, say, when it is in the complement of *for*), then the ECM object should cause LOCAL to be updated, rendering the matrix subject LOCAL:0 and outside the scope of reflexive reference, in which case sentences like (122c) should be ungrammatical. At present I do not see an obvious solution for this problem, and so set the issue aside for future work.

(122) a. Hannibal$_i$ believes himself$_i$ to have been purged of lesser rudenesses.

b. Will believes Hannibal$_i$ to have purged himself$_i$ of lesser rudenesses.

c. Will would prefer for himself to avoid becoming Hannibal's "murder husband".

In any event, this discussion has highlighted the key weaknesses inherent in cue-based approaches to explaining linguistic dependencies: the theory we adopt is only as good as the cues we posit, and the mechanism by which those features are valued is itself a non-trivially difficult puzzle. Kush's model gets closest of any current account to providing an algorithmic way of identifying local, c-commanding antecedents, and even it struggles to encode some of the more nuanced cases discussed in syntactic Binding Theories. Thus, we have reason to remain somewhat dissatisfied with the model proposed in this dissertation. While cue-based retrieval provides a powerful, domain-general means of describing the on-line resolution of linguistic dependencies, its ability to encode the linguistic constraints of interest (i.e. Binding Theory) remains only partially understood, and in need of further development. With respect to the proposed model of logophlexivity, this would entail a more thorough investigation of mechanisms for encoding locality and c-command, perhaps expanding on Kush (2013)'s algorithm. In addition, the mechanism by which referents are mapped to the discourse role hierarchy needs further formalization to be integrated more thoroughly with the ACT-R framework.

However, even allowing for these difficulties, it would be premature to reject cue-based models wholesale. These models are extremely general, and allow for the description of effects at both linguistic, and general cognitive levels of explanation. Given that the linguistic system must, at some level, be situated in the more general domain of greater cognition, trying to locate linguistic processing within a unified model of memory seems desireable, if not necessary. Moreover, as noted above, at least some of the explanatory power for the logophlexives model is derived from the retrieval mechanism itself. In explaining the target-match asymmetry, I critically rely on the link between cue-matching and probability of re-activation inherent in the retrieval mechanism. Consequently, the difficulties discussed here should not be seen as reasons to reject a cue-based

approach to reflexive antecedent identification, but rather a call to more closely consider the representations over which such models must operate and the means by which they can be encoded.

5.3.6 Summary: logophlexivity and reflexive comprehension

In this section, I have attempted to tie the logophlexives model back into the literature on real-time reflexive comprehension, in addition to considering whether logophlexivity is a sub-case of grammatical illusion. Overall, the logophlexives model seems to fare well with previous findings in the reflexives comprehension literature, correctly predicting the distribution of sensitivity to non-local referents across several experiments. However, there were two cases which did not conform to the model's predictions. First, Parker and Phillips (2017)'s Experiment 2 presented a case sensitivity to lures which were not, on their face, particularly good logophoric antecedents. However, the structures employed in this experiment resembled cases of sub-command in Mandarin, suggesting that an alternative grammatical explanation may yet be extended to these data. The second non-conforming datum was seen in King et al. (2012), whose data remain mysterious, and will need to be the focus of much future work. Finally, I argued that logophlexivity should not be grouped into the class of grammatical illusions, and that the nature of the retrieval mechanism, while integral to the model, is not the primary explanatory force behind logophlexivity. Notably absent from this discussion was any mention of the processing signature of non-argument reflexives. This discussion is found in the following section, where it finds a more natural home in our re-evaluation of predicate-based theories of binding.

5.4 Logophlexivity and Binding Theory

This penultimate section of Chapter 5 revisits the various models of binding considered in Chapter 1 and considers various implications presented by the logophlexives model. Primarily, this will constitute a reevaluation of predicate based models on the basis that a division between argument and non-argument reflexives may not be necessary to derive the facts they are concerned with. First, however, it is helpful to remind ourselves of the relation of logophlexivity to the Standard Binding Theory of Chomsky (1986). Because of the position of OP_{log} in the model, the logophlexives hypothesis is actually entirely consistent with the SBT model. That is, under the current theory, all instances of reflexive binding are local binding, even those with apparently long distance reference. Thus, the only extension needed to accommodate long-distance reference in the standard framework is the assumption of a logophoric operator capable of binding animate reflexives. A similar position has already been advanced in Charnavel and Sportiche (2016). Given

this, the question is whether making a distinction between argument and non-arguments is still necessary. In what follows, I argue that although argument and non-argument reflexives may yet show different behavior, these differences should not be understood in terms of a predicate-based version of Binding Theory.

5.4.1 Reconsidering predicate-based models binding

As seen in Chapter 1, predicate-based models are aimed at capturing three facts: (1) the strong intuitions that argument reflexives are obligatorily locally bound, but that non-argument reflexives may refer non-locally (2) the overlap in distribution of pronominal and reflexive referring devices; (3) the distribution of intermediate forms like *zich* (Pollard & Sag, 1992; Reinhart & Reuland, 1993). Each of these observations poses a challenge for the SBT model, and so they should, ideally, be subsumed within the new logophlexives hypothesis. I address each point in turn.

First, the results of Experiments 1-5 show that the empirical generalization in point (1) does not entirely hold. In all five experiments, comprehenders demonstrated substantial sensitivity to non-local referents while processing, judging, and interpreting embedded, direct-object reflexives. Given that these reflexives were co-argument with other referents, these results are unexpected under predicate-based binding models. All the same, the impression that non-local reference is more difficult for argument reflexives persists, and thus deserves an explanation. As before, one could appeal here to the prosodic constraints noted by (Charnavel & Sportiche, 2016), and by Ahn (2015). Logophoric interpretations of reflexives seem to be preferentially associated with nuclear stress on the reflexive, which may be instrumental in licensing their realization. If the appropriate stress is easier to realize for non-argument reflexives than for argument reflexives, then at least some of the intuition is explained[16]. Beyond this, it may be possible to extend the findings of King et al. (2012) to account for less stringent locality effects for non-argument reflexives. Under this view, argument reflexives are strictly locally bound because of their adjacency to verbs. This adjacency results in the co-activation of the reflexive and any pre-verbal arguments which were retrieved for thematic integration. It is this coactivation, then, which leads comprehenders to strongly prefer the local interpretation. In contrast, non-argument reflexives are typically not verb-adjacent, perhaps allowing for activation decay of verbal arguments, and granting the reflexive slightly more freedom of reference.

[16]See also Charnavel and Sportiche (2016)'s appeal to Cardinaletti and Starke (1999)'s generalization regarding the choice of referring device given a strong/weak distinction

Interestingly, this latter proposal suggests a novel interpretation of the facts for possessed NP embedded reflexives. Recall that all three binding models had to posit that possessors acted as subjects within the DP. For (Chomsky, 1986), this had the result of making a possessive DP the locality domain for an embedded reflexive, while for (Reinhart & Reuland, 1993) and (Pollard & Sag, 1992) this meant that the reflexive had a co-argument, and therefore was obligatorily bound. Under the view that co-argumenthood restrictions are (at least partially) due to the strong activation of local thematic DPS, we may have a fairly different interpretation of these facts. Rather than restricting the size of the locality domain, or imposing obligatory co-argument binding, possessors may simply be more accessible for reference due to their recency and high base-activation. If so, the SBT model could actually be simplified such that relativizing the size of the locality domain is no longer necessary. It is somewhat less clear how this explanation might affect predicate-based models, as arguments for these models relies, in part, on the observation that possessed NP reflexives *must* be locally bound. Relevant to this point, Runner, Sussman, and Tanenhaus (2003, 2006) actually demonstrated that comprehenders *do* adopt interpretations in which the reflexive does not refer to the possessor, indicating that if this is a grammatical constraint, it is not uniformly enforced. However, if we suppose that resting activation, rather than grammatical principles, is responsible for a possessor-preference, then this finding is unsurprising—a cue-based implementation of antecedent identification predicts that, some proportion of the time, a non-possessor antecedent will be adopted.

Moreover, this account makes the interesting prediction that simply interpolating (temporal) distance between a verb and a reflexive should collapse the distinction between argument/non-argument anaphors, rendering them equally likely to take a non-local referent. As noted above, tentative evidence for this position comes from (King et al., 2012), who found lure-match facilitation when a verb and reflexive were not adjacent. However, their manipulation did not provide a comparison of argument and non-argument reflexives, and so more work is needed to test this possibility. At present, examples like (123) and 124 provide a promising potential contrast for future investigations. To the extent that (123b) is on-par with the two examples in (124), we would have evidence that distance from the verb, rather than co-argumenethood impacts the likelihood of non-local reference.

(123) a. $*$ John said that Mary emailed himself a picture of Jill.

b. ? John said that Mary emailed a picture of Jill to himself.

(124) a. John said that Mary emailed a picture of himself to Jill.

b. John said that Mary emailed Jill a picture of himself.

The second datum to be explained is the overlap in distribution of reflexive and pronominal forms, particularly in Picture NP reflexives. Handily, Kush (2013)'s implementation of the LOCAL feature already captures this fact for us. Recall that the overlap of referring devices was exactly his concern in sentences like (121). To explain this, he relativized the valuation of LOCAL for pronouns to their position in the sentence, such that pronouns not on the clausal spine employ a retrieval probe underspecified for LOCAL. However, this seems to simply be a procedural implementation of non-argumenthood, as argument referring devices will, for the most part, exist on the clausal spine. Thus, while the logophlexives model *can* accommodate overlap among referring devices, it perhaps does so by sneaking a distinction between argument and non-argument positions into the manner in which local-cued retrievals are initiated. Thus, if we are to maintain no argument/non-argument distinction, there seems to be a problem with the way LOCAL feature valuation happens. Intuitively, we would like to recapitulate Chomsky (1986)'s definition of local, such that DPs constitute the local domain for pronouns, but not for reflexives, but this isn't quite how LOCAL is defined. At present, I know of no solution to this problem, but further work on the appropriate definition of LOCAL is needed if we are to entirely do away with the argument/non-argument distinction. As it currently stands, the logophlexives hypothesis presents a kind of hybrid model: the fundamental explanation for non-local reference does not derive from co-argumenthood, but overlap in the distribution of referring devices does.

Finally, the logophlexives model has relatively little to say about referring devices of intermediate status like *zich*. Here, the core challenge lies in deriving the fact that *zichzelf* must always be locally bound, but *zich* may only be locally bound under certain circumstances (e.g. in the complements of predicative prepositions, inherently reflexive predicates, or ditransitives). One possibility would be to suppose that Dutch has actually split its pronominal system such that *zich may* be bound by OP_{log}, while *zichzelf* may not be. If so, we could explain the binding facts for *zich* by allowing that the subjects of inherently reflexive predicates and predicative PPs[17] are more likely to control high positions on Sells' hierarchy. In this case, these subjects would be likely to co-refer with OP_{log}, and therefore good (indirect) antecedents for *zich*. Moreover, this analysis provides a possible explanation of the dispreference for *zichzelf* in these constructions. If *zichzelf* is, in some sense, anti-logophoric, then binding by an antecedent which itself is co-referent

[17]Playing slightly fast and loose with the notion of "subject", here, so that the subject of the predicate *modified* by the predicative PP is the PP's subject.

with OP_{log} may be dispreferred. That said, it isn't clear how to extend this analysis to ditransitives, as *zich* may occupy either object role, as long as the other is *zichzelf*, and multiple uses of *zichzelf* is degraded. Thus, the dutch pronominal system remains something of a puzzle under the logophlexives hypothesis, and perhaps the strongest data in favor of a predicate-based binding model.

With regards to Binding Theory, then, the model and evidence presented here are more closely aligned with SBT, and related theories. In particular the the results of Experiments 1-5 present a strong empirical challenge for predicate-based theories of binding, since they demonstrate that non-local reference is not dependent on non-argumenthood. Accordingly, the logophlexives model treats non-local reference as a matter of local binding, in the manner of Charnavel and Sportiche (2016). That said, the argument/non-argument distinction still seems to be necessary to account for the overlapping distribution of reflexives and pronominals in non-argument positions, as seen in the current formulation of the retrieval cue LOCAL. Further work on the algorithm for valuing this feature may render this distinction unnecessary, but this solution is not presently available. Finally, the strongest evidence in favor of predicate based models is found in the distribution of *zich* in Dutch, which cannot currently be subsumed into the model of logophoricity presented here. Given this somewhat mixed picture, it may actually be blessing that the logophlexives model is something of a hybrid, combining the locality restrictions of the SBT model with the argument/non-argument distinction of predicate-based theories.

5.4.2 Reconsidering the processing of exempt anaphora

Given the discussion above, we can begin to better understand the sentence processing literature on exempt anaphora. Before beginning, it is important to note that many of the studies reported here reference "logophoric processing" in their discussions. However, this is a different meaning of the word "logophor" than I have been using in this dissertation. Where I have intended "logophor" to mean "pronouns which take OP_{log} as an antecedent". In contrast, these studies adopt Reinhart and Reuland (1993)'s use of the word to mean "non-argument reflexives", regardless of whether these reflexives are taking a logophoric center as an antecedent. Consequently, these studies have relatively little to say about the process by which OP_{log} is selected for reference, but can be informative about how the processing of argument and non-argument reflexives may differ.

There are three primary findings in the literature on exempt reflexives in need of explanation. First, studies like Piñango and Burkhardt (2005); Burkhardt (2005); Harris, Wexler, and Holcomb

(2000); Cunnings and Sturt (2014) report that interpreting reflexives in exempt positions is more difficult than interpreting those in non-exempt positions. This fact might be accommodated in the logophlexives model by supposing that OP_{log} is easier to access for non-argument reflexives, as suggested in the discussion above. If so, it may be that OP_{log} and local DPs are in greater competition with each other in non-argument positions, possibly leading to greater comprehension difficulty. This would be analogous to the "multiple match" effect reported by (Badecker & Straub, 2002), and given that we don't observe multiple match effects in any of our studies, this solution seems at first implausible. However one critical difference here is that non-argument reflexives (assuming they are not in possessed DPs) will be relatively distant from the verb, possibly allowing for the thematic arguments to have decayed by the time they are encountered. If so, then OP_{log} and the local DPs may be in greater competition than they otherwise would be, and may give rise to a kind of multiple match effect. Alternatively, it may be that non-argument reflexives really are preferentially accessing OP_{log}, and that resolving the reference of OP_{log} occurs at a delay. At present, I cannot arbitrate between these alternatives.

The second finding for exempt anaphors is that comprehenders preferentially attend to the information SOURCE when interpreting them (Kaiser et al., 2009). In (Kaiser et al., 2009)'s study, reflexives embedded in picture-noun-phrases were interpreted as referring to the subject of the verb *tell* more often than the subject of the verb *hear*. This is straightforwardly compatible with the logophlexives model, under which *tell* subjects will be probablistically mapped to SOURCE, and therefore likely to control OP_{log}, while subjects of *hear* will, at, best, control PIVOT, and relatively less likely to control OP_{log}. The fact that the reflexives is embedded in an NP is, from the logophlexives model's perspective, irrelevant.

The final point is actually a contradiction, of sorts: (*i*) as noted above, comprehenders do not obligatorily take possessors as the antecedents of reflexives embedded in possessed noun phrases (Kaiser et al., 2009; Runner et al., 2006, 2003); (*ii*) even reflexives in exempt position tend to be robust to the influence of non-local referents (Cunnings & Sturt, 2014). Again, the solution here lies in the referential properties of OP_{log}. First, as noted above, if we abandon the notion that np-embedded reflexives are obligatorily bound by their possessors, then we can derive a strong preference for possessor binding from its relatively high base activation. Given that retrieval is probablistic, we then exepect that a non-possessor antecedent will be selected some percentage of the time. Interestingly, (Kaiser et al., 2009) report more attention to non-possessors which are the subjects of speech verbs, congruent with the idea that comprehenders are attending to OP_{log} as an alternative antecedent to the possessor.

To explain Cunnings and Sturt (2014)'s findings, then, recall that non-local referents in their study were predominately not the subjects of speech predicates. Once again, then, the logophlexives model predicts that these referents should be relatively less likely to co-refer with OP_{log}, and therefore less available as antecedents for an embedded reflexive. Thus, it may be that the critical difference between Cunnings and Sturt (2014) and Kaiser et al. (2009) lies in the choice of verbs used, as suggested at the outset of Chapter 2. The logophlexives hypothesis thus resolves the tension between these two studies by suggesting that verb-type influences the likelihood of where on Sells' hierarchy a given referent will be mapped.

Given this discussion, while it does seem that there are differences between the processing of argument and non-argument reflexives, there is not strong evidence that this difference is grammatical in nature. That is, we can explain variation in binding by possessors, and preferences for reference to sources, without appealing to the syntactic position of the reflexive in question. Instead, these facts fall out from a combination of antecedent recency (i.e. base activation) and the referential properties of OP_{log}. Likewise, evidence that non-argument reflexives are more difficult to process may receive an explanation which does not require Binding Theory being relativized to different syntactic configurations. The findings from the sentence processing literature are thus fairly consistent with the discussion of Binding Theory models above: argument and non-argument reflexives *do* behave differently, but it isn't clear that this is due to a predicate-based version of Binding Theory.

5.4.3 Summary: logophlexivity and Binding Theory

In this section I have presented an (admittedly incomplete) argument against predicate-based Binding Theories. The core of this argument is that, empirically, non-local reference is *not* confined to non-argument reflexives, a fact predicted by an OP_{log}-augmented SBT model, but unexpected under any predicate based theory. The remainder of the chapter then struggled to understand why non-argument reflexives seem to behave so differently, if Binding Theory itself isn't responsible. Here I appealed to prosodic and contextual constraints which might disfavor OP_{log} as a binder for argument reflexives (Charnavel & Sportiche, 2016; Ahn, 2015). In addition, I extended King et al. (2012)'s hypothesis to suggest that co-activation of thematic arguments may be responsible for the strong local-binding preference of direct object reflexives. This possibility was particularly important in explaining why non-argument reflexives might be more difficult to process than their argument counterparts (Piñango & Burkhardt, 2005; Harris et al., 2000; Cunnings & Sturt, 2014). Taken together, these elements present a strong argument against the need for incorporating argu-

menthood into our models of binding—co-argumenthood does not cleanly predict the availability of (non-)local reference, and to the extent that it does affect reflexive comprehension it might be explainable in terms of independent factors.

However, the argument remains incomplete. Two aspects of predicate based theories remain to trouble us. First, to explain the overlap of pronominal and reflexive reference, the logophlexives model had to covertly make use of the argument/non-argument distinction in its use of the feature LOCAL to search for an antecedent. Second, and perhaps more troublingly, the model currently cannot assign a unified analysis to the Dutch pronominal system. In particular the division of labor between *zich* and *zichzelf* in ditransitives remains mysterious, and in need of future investigation. Consequently, while the data and model presented here are more conceptually consistent with the SBT model, they do not decisively rule out an approach to Binding Theory centered around predicates and co-arguments.

5.5 Stray observations and future directions

In this chapter I have attempted to bring together the various components of the dissertation into a unified discussion of the evidence for, and representation of, logophoricity in English reflexive comprehension. The primary evidence came from experiments demonstrating that factors which impact the likelihood of perspective assignment directly impact the availability of non-local antecedents for reflexive pronouns. Based on this evidence, I proposed the "logophlexives model", by which the retrieval operation engaged by a reflexive may occasionally select OP_{log}, a silent logophoric operator in the left periphery of embedded clauses which refers to the highest specified discourse role on Sells (1987)'s scale. The critical effects of Experiments 1-5 were then explained by the ϕ-deficient nature of OP_{log}, as well as the probability with which different kinds of referents were mapped on to Sells' scale. We then saw that this model can account for substantial variability in previous findings (Sturt, 2003b; Xiang et al., 2009; Dillon et al., 2013; Parker & Phillips, 2017; Kaiser et al., 2009; Cunnings & Sturt, 2016), though a handful of points remain mysterious (King et al., 2012).

The remainder of the chapter was taken up with a consideration of the role played by retrieval mechanisms in the logophlexives model, as well as this model's implications for Binding Theory. With respect to the former point, I argued that while cue-based frameworks were appropriate as a means of implementing the logophlexives model, much work remains to be done in specifying precisely how linguistic constraints are encoded within such a system. Moreover, we saw that

the retrieval mechanism itself was not the critically explanatory part of the model. Instead, the explanations proposed here find their root in procedural algorithms often taken for granted in retrieval modeling. Lastly, I advanced an argument against predicate-based Binding Theories, on the basis that non-local reference does not appear to be conditioned on non-argumenthood.

Looking towards future work, the discussion here has presented several possible avenues of investigation. First, given that OP$_{log}$ lacks ϕ features, I speculated that checking antecedent/reflexive congruence when the referent is OP$_{log}$ should be more difficult, as the comprehender must consult their discourse model. At present, there is insufficient data to address this question—none of the studies in Experiments 1-5 were aimed at addressing it, and evidence of "logophoric difficulty" (e.g. Piñango & Burkhardt, 2005; Burkhardt, 2005) is actually an example of difficulty with exempt anaphora. Future work, then, will need to address this question more carefully.

Another potential point of investigation derives from the model's characterization of referent assignment to Sells' hierarchy. In the formulation given here, indexical pronouns should be preferentially mapped higher on Sells' hierarchy than the subjects of perception verbs. Given this, we should expect person blocking effects to be even more dramatic when the matrix subject is the subject of a perception verb. This should be a fairly straightforward modification of the materials used in Experiments 3 an 4, and as such is an obvious starting point for future work.

One interesting prediction not discussed above is found in the fact that OP$_{log}$ is ϕ deficient, a fact I used to explain the target-match asymmetry. However, this also predicts that if a reflexive antecedent search is *not* cued with morphosyntax, OP$_{log}$ should be a relatively more competitive antecedent, and long-distance reference should be achieved more easily. Descriptively, this seems to be the case. The three "long-distance reflexive" languages discussed in this dissertation (Mandarin, Japanese, and Icelandic) all use reflexive forms which do not inflect for number, gender, or person (*ziji*, *zibun*, *sig*)[18]. Thus, the prediction from ϕ-deficient OP$_{log}$ seems to be (tentatively) borne out cross-linguistically, though further investigation is necessary to see if the pattern holds.

The final experimental prediction of this section is derived from the hypothesis put forward by King et al. (2012). Much of the argument in Section 5.4 revolved around the assumption that recently activated referents will act as strong attractors for reference, and that this fact can be used to explain differences among (non-)argument reflexives. A tentative paradigm for testing this prediction was given above, and this seems like a very promising avenue for future experiments.

[18]However, Icelandic *sig* is obligatorily third person, and there exists a possessive reflexive form which does inflect for gender and number. Icelandic may also be problematic in that it may not display person blocking effects.

Finally, as the discussion in Section 5.3.2 made clear, substantially more work needs to be conducted to establish the mechanisms by which linguistic constraints are encoded in a cue-based retrieval framework. Models articulated in these systems are only as adequate as the cues they posit. In the absence of a model of how those features are identified, valued, and maintained, the theory can only rely on error-prone retrieval as the explanatory mechanism. And as I hope this dissertation has convinced you, that mechanism simply cannot extend to all puzzles.

APPENDIX

STATISTICAL MODELING RESULTS

Table A.1. Experiment 1a: Fixed effect coefficients (standard error in parentheses) for linear regression models fit to sentence ratings in Experiment 1a. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	4.35 (0.10)	43.30
TARGET	**0.83 (0.08)**	10.23
LURE	0.02 (0.03)	0.88
VERB	-0.01 (0.02)	0.54
TARGET×LURE	**-0.05 (0.03)**	2.15
TARGET×VERB	-0.04 (0.03)	1.22
LURE×VERB	0.04 (0.02)	1.89
TARGET×LURE×VERB	-0.02 (0.03)	0.58

Table A.2. Experiment 1b: Mixed Effects model coefficients and standard errors for fixation-duration measures. Effects which were significant ($|t| > 2$) are given in bold-face

	Reflexive			Spillover		
Fixed Effects	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
INTERCEPT	341 (11)	513 (26)	585 (26)	419 (19)	651 (34)	705 (37)
TARGET	6 (6)	**55 (17)**	**66 (13)**	0 (8)	97 (19)	**47 (13)**
LURE	4 (5)	23 (13)	14 (9)	-7 (5)	10 (15)	-2 (11)
VERB	-4 (4)	1 (13)	-10 (10)	-12 (6)	29 (21)	-11 (11)
TARGET×LURE	-2 (4)	**29 (14)**	18 (10)	-7 (5)	**31 (15)**	-3 (10)
TARGET×VERB	-7 (4)	4 (11)	8 (9)	7 (6)	31 (16)	8 (10)
LURE×VERB	-7 (4)	-21 (15)	-11 (8)	-4 (7)	19 (15)	-4 (12)
TARGET×LURE×VERB	-7 (4)	**-38 (15)**	**-20 (8)**	8 (5)	8 (17)	-6 (13)

Table A.3. Experiment 1c: Fixed effect coefficients (standard error in parentheses) for logistic regression model fit to proportion matrix responses in Experiment 1c. Significant effects ($p<.05$) are given in bold-face

Fixed Effects	$\hat{\beta}$	z	p
INTERCEPT	1.42 (0.18)	7.87	$<.001$
LURE	**-0.34 (0.09)**	-3.74	$<.001$
VERB	**-0.24 (0.09)**	-2.72	$<.01$
LURE×VERB	0.00 (0.08)	0.04	0.97

Table A.4. Experiment 2a: Fixed effect coefficients (standard error in parentheses) for linear regression models fit to sentence ratings. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	4.30 (0.12)	35.32
TARGET	**0.16 (0.04)**	3.63
LURE	**0.89 (0.14)**	6.35
VERB	0.07 (0.04)	1.72
TARGET×LURE	**-0.24 (0.06)**	-3.77
LURE×VERB	-0.05 (0.04)	-1.38

Table A.5. Experiment 2b: Mixed Effects model coefficients and standard errors for fixation-duration measures. Effects which were significant ($|t| > 2$) are given in bold-face

	Reflexive			Spillover		
Fixed Effects	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
INTERCEPT	300 (11)	411 (22)	437 (17)	383 (19)	582 (43)	547 (28)
TARGET	2 (6)	30 (15)	**28 (13)**	14 (7)	**104 (25)**	**48 (16)**
LURE	5 (8)	-21 (22)	**42 (17)**	8 (10)	**115 (38)**	**47 (18)**
VERB	-4 (5)	-24 (16)	-10 (10)	9 (7)	-6 (26)	-5 (11)
TARGET×LURE	12 (11)	9 (31)	**50 (23)**	30 (17)	**135 (53)**	**91 (25)**
LURE×VERB	3 (10)	4 (38)	-1 (19)	-10 (17)	-29 (61)	-27 (24)

Table A.6. Experiment 3a: Fixed effect coefficients (standard error in parentheses) for linear regression models fit to sentence ratings. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	3.89 (0.15)	26.3
TARGET	**1.29 (0.15)**	8.35
LURE	**0.16 (0.04)**	4.34
PERSON	**0.23 (0.08)**	3.01
TARGET×LURE	**-0.21 (0.05)**	-4.61
LURE×PERSON	0.07 (0.06)	1.32

Table A.7. Experiment 3b: Mixed effect model coefficients and standard errors for sentence naturalness ratings in Experiment 3b. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	4.10 (0.15)	27.59
TARGET	**1.40 (0.15)**	9.42
LURE	**0.20 (0.04)**	5.42
PERSON	**0.19 (0.04)**	4.2
TARGET×LURE	**-0.23 (0.06)**	-4.16
LURE×PERSON	**0.12 (0.05)**	2.55

Table A.8. Experiment 3c: Mixed Effects model coefficients and standard errors for fixation-duration measures. Effects which were significant ($|t| > 2$) are given in bold-face

	Reflexive			Spillover		
Fixed Effects	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
INTERCEPT	311 (13)	416 (21)	496 (28)	386 (20)	659 (46)	608 (42)
TARGET	10 (6)	**52 (16)**	**82 (13)**	10 (11)	**172 (29)**	**87 (17)**
LURE	**10 (4)**	7 (11)	**33 (9)**	6 (6)	56 (28)	**32 (15)**
PERSON	**16 (5)**	29 (15)	**50 (13)**	**16 (7)**	43 (30)	**57 (15)**
TARGET×LURE	**19 (6)**	5 (15)	**38 (16)**	6 (8)	38 (32)	18 (15)
LURE×PERSON	-3 (5)	**-47 (13)**	-7 (11)	-1 (8)	51 (31)	9 (13)

Table A.9. Experiment 4a: Mixed effect model coefficients and standard errors for sentence naturalness ratings in Experiment 4a. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	3.92 (0.11)	34.94
TARGET	**1.36 (0.13)**	10.28
LURE	**0.10 (0.03)**	3.36
PERSON	**0.12 (0.04)**	2.78
TARGET×LURE	**-0.13 (0.05)**	-2.82
LURE×PERSON	0.04 (0.04)	0.95

Table A.10. Experiment 4b: Mixed Effects model coefficients and standard errors for fixation-duration measures. Effects which were significant ($|t| > 2$) are given in bold-face

	Reflexive			Spillover		
Fixed Effects	*First Pass*	*Go-Past*	*Total Time*	*First Pass*	*Go-Past*	*Total Time*
INTERCEPT	293 (9)	392 (18)	483 (21)	382 (19)	621 (37)	596 (35)
TARGET	3 (6)	**41 (13)**	**79 (14)**	4 (8)	**139 (22)**	**66 (15)**
LURE	**8 (4)**	-2 (8)	12 (9)	-8 (7)	**33 (15)**	5 (10)
PERSON	6 (6)	9 (17)	**54 (18)**	0 (9)	**66 (24)**	**50 (14)**
TARGET×LURE	0 (6)	10 (12)	21 (13)	4 (8)	**69 (21)**	27 (13)
LURE×PERSON	0 (6)	12 (17)	-16 (13)	-8 (8)	**-57 (21)**	-31 (17)

Table A.11. Meta Analysis: Fixed effect coefficients (standard error in parentheses) for linear regression models fit to pooled data from Experiments 3a, 3b and 4a, collapsing across first and second person pronouns ("Indexical"). Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	3.83 (0.08)	48.40
TARGET	**1.48 (0.03)**	45.48
INDEXICAL	**-0.23 (0.03)**	6.85
ANIMACY	-0.08 (0.04)	1.81
LURE	**0.18 (0.02)**	8.27
TARGET×LURE	**-0.21 (0.03)**	6.71
INDEXICAL×LURE	**-0.08 (0.03)**	2.51
ANIMACY×LURE	**-0.10 (0.04)**	2.68

Table A.12. Experiment 5: Fixed effect coefficients (standard error in parentheses) for linear regression models fit to sentence ratings in Experiment 5. Significant effects ($|t| \geq 2$) are given in bold-face

Fixed Effects	$\hat{\beta}$	t
INTERCEPT	4.59 (0.09)	51.87
TARGET	**1.08 (0.10)**	11.01
LURE	0.03 (0.04)	0.68
REFLEXIVE	0.00 (0.05)	0.01
TARGET×LURE	-0.03 (0.03)	0.79
TARGET×REFLEXIVE	0.07 (0.04)	1.76
LURE×REFLEXIVE	0.00 (0.03)	0.11
TARGET×LURE×REFLEXIVE	-0.04 (0.03)	1.43

REFERENCES

Ahn, B. (2015). *In progress. giving reflexivity a voice: Twin reflexives in english* (Unpublished doctoral dissertation). Doctoral Dissertation, UCLA.

Anand, P. (2006). *De de se (doctoral dissertation).* MIT.

Andrews, C., Yacovone, A., Sloggett, S., & Dillon, B. (2016). *Reflexives: We don't see the attraction.* Poster at AMLaP 22: Bilbao, Spain.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 748.

Burkhardt, P. (2005). *The syntax-discourse interface: Representing and interpreting dependency* (Vol. 80). John Benjamins Publishing.

Cardinaletti, A., & Starke, M. (1999). The typology of structural deficiency: A case study of the three classes of pronouns. *Clitics in the languages of Europe, Mouton de Gruyter, Berlin*, 145–233.

Charnavel, I., & Sportiche, D. (2016). Anaphor binding: What french inanimate anaphors show. *Linguistic Inquiry*.

Charnavel, I., & Zlogar. (2015). English reflexive logophors.

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? evidence from mandarin chinese. *Empirical approaches to linguistic theory: Studies in meaning and structure*, *111*, 43.

Chomsky, N. (1981). *Lectures on government and binding: The pisa lectures* (No. 9). Walter de Gruyter.

Chomsky, N. (1986). *Barriers*. MIT press.

Clements, G. N. (1975). The logophoric pronoun in ewe: Its role in discourse. *Journal of West African Languages*, *10*, 141-177.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. *Eye movements: A window on mind and brain*, *27*, 341–72.

Cole, P., & Wang, C. (1996). Antecedents and blockers of long-distance reflexives: The case of chinese ziji. *Linguistic inquiry*, 357–390.

Culy, C. (1994). Aspects of logophoric marking. *Linguistics*, *32*(6), 1055–1094.

Culy, C. (1997). Logophoric pronouns and point of view. *Linguistics*, *35*(5), 845–860.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*(1-2), 188–219.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

Cunnings, I., & Sturt, P. (2016). Locality and discourse constraints in reflexive resolution: Evidence from eyemovements during reading. In *Poster at amlap 22: Bilbao, spain.*

Dillon, B. (2014). Syntactic memory in the comprehension of reflexive dependencies: an overview. *Language and Linguistics Compass*, *8*(5), 171–187.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103.

Dillon, B., Staub, A., Levy, J., & Clifton Jr, C. (2017). Which noun phrases is the verb supposed to agree with?: Object agreement in american english. *Language*, *93*(1), 65–96.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: number agreement in sentence production. *Psychological Review*, *112*(3), 531.

Engelmann, F., Jäger, L. A., & Vasishth, S. (Submitted). The determinants of retrieval interference in dependency resolution: Review and computational modeling. *Journal of Memory and Language*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models* (Vol. 1). Cambridge University Press Cambridge.

Grant, M., Dillon, B., & Sloggett, S. (2015). *Similarities in processing attachment and pronominal ambiguities.* Poster at CUNY 28: Los Angeles, California.

Hammerly, C., & Dillon, B. (2017). *Restricting domains of retrieval: Evidence for clause-bound processing from agreement attraction.* Poster at CUNY 30: Cambridte, MA.

Harris, T., Wexler, K., & Holcomb, P. (2000). An erp investigation of binding and coreference. *Brain and Language*, *75*(3), 313–346.

He, X., & Kaiser, E. (2012). Is there a difference between 'you'and 'i'? a psycholinguistic investigation of the chinese reflexive ziji. *University of Pennsylvania Working Papers in Linguistics*, *18*(1), 12.

Heim, I. (2001). *Semantics and morphology of person and logophoricity.* (Talk given at the University of Tübingen)

Huang, C.-T. J., & Liu, C.-S. L. (2001). Logophoricity, attitudes, and ziji at the interface. *Long-distance reflexives*, 141–195.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, *112*(1), 55–80.

King, J., Andrews, C., & Wagers, M. (2012). Do reflexives always find a grammatical antecedent for themselves. In *25th annual cuny conference on human sentence processing*.

Koopman, H., & Sportiche, D. (1989). Pronouns, logical variables, and logophoricity in abe. *Linguistic Inquiry*, 555–588.

Kreiner, H., Sturt, P., & Garrod, S. (2008). Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements. *Journal of Memory and Language*, *58*(2), 239–261.

Kuno, S. (1972). Pronominalization, reflexivization, and direct discourse. *Linguistic Inquiry*, *3*(2), 161–195.

Kuno, S. (1986). Anaphora in japanese. In *first sdf workshop in japanese syntax* (pp. 11–70).

Kush, D. W. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (Unpublished doctoral dissertation). University of Maryland College Park.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, *29*(3), 375–419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, *10*(10), 447–454.

Loss, S. (2014). Iron range english reflexive pronouns. In R. Zanuttini & L. R. Horn (Eds.), *Micro-syntactic variation in north american english.* Oxford University Press, USA.

Maling, J. (1984). Non-clause-bounded reflexives in modern icelandic. *Linguistics and Philosophy*, *7*(3), 211–241.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, *29*(2), 111–123.

McElree, B. (2006). Accessing recent events. *Psychology of learning and motivation*, *46*, 155.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research*, *18*(1), 5–19.

Parker, D. (2014). *The cognitive basis for encoding and navigating linguistic structure* (Unpublished doctoral dissertation). University of Maryland College Park.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321–339.

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

Patil, U., Vasishth, S., & Lewis, R. (2016). Retrieval interference in syntactic processing: The case of reflexive binding in english. *Frontiers in Psychology*, *7*, 329.

Pearson, H. (2015). The interpretation of the logophoric pronoun in ewe. *Natural Language Semantics*, *23*(2), 77–118.

Piñango, M. M., & Burkhardt, P. (2005). Pronominal interpretation and the syntax-discourse interface: Real-time comprehension and neurological properties. *Anaphora processing: Linguistic, cognitive and computational models*, 221–238.

Pollard, C., & Sag, I. A. (1992). Anaphors in english and the scope of binding theory. *Linguistic inquiry*, *23*(2), 261–303.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic inquiry*, *24*(4), 657–720.

Rizzi, L. (1990). *Relativized minimality.* The MIT Press.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements. *Cognition*, *89*(1), B1–B13.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2006). Processing reflexives and pronouns in picture noun phrase. *Cognitive Science*, *30*(2), 193–241.

Schlueter, Z., Parker, D., & Lau, E. F. (2017). *(mis)interpreting agreement attraction: Evidence from a novel dual-task paradigm.* Talk at CUNY 30: Cambridge, MA.

Sells, P. (1987). Aspects of logophoricity. *Linguistic Inquiry*, *18*(3), 445–479.

Speas, M., & Tenny, C. (2003). Configurational properties of point of view roles. *Asymmetry in grammar*, *1*, 315–345.

Sturt, P. (2003a). A new look at the syntax-discourse interface: The use of binding principles in sentence processing. *Journal of psycholinguistic research*, *32*(2), 125–139.

Sturt, P. (2003b). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*(3), 542–562.

Sturt, P. (2013). Syntactic constraints on referential processing. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 136–159). Psychology Press East Sussex.

Sturt, P., & Kwon, N. (2017). *Agreement attraction: roles of active dependencies and attractor position.* Poster at CUNY 30: Cambridge, MA.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Tang, C.-C. J. (1989). Chinese reflexives. *Natural Language & Linguistic Theory*, *7*(1), 93–121.

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 407.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of memory and language*, *65*(3), 247–263.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.

von Stechow, A. (2002). *Binding by verbs: tense, person and mood under attitudes.* (unpublished manuscript, University of Tübingen)

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: Erp evidence. *Brain and Language*, *108*(1), 40–55.

Xiang, M., Grove, J., & Giannakidou, A. (2013). Dependency-dependent interference: Npi interference, agreement attraction, and global pragmatic inferences. *Frontiers in psychology*, *4*.

Xue, P., Pollard, C., & Sag, I. A. (1994). A new perspective on chinese ziji. In *the proceedings of the thirteenth west coast conference on formal linguistics.*