

Recoverability Optimality Theory: Discourse Anaphora in a Bidirectional framework

Adam Buchwald, Oren Schwartz, Amanda Seidl, and Paul Smolensky*

The Johns Hopkins University
Department of Cognitive Science
243 Krieger Hall
3400 N. Charles St. Baltimore, MD 21218-2685, U.S.A

Abstract

In this work we explore the possibility of formalizing explanations based in the notion “deletion under recoverability”: information should be omitted in expressions if that information can be recovered without being overtly expressed – e.g., from information already encoded in the state of the discourse. This style of explanation is difficult to formalize for many reasons, among them that the contradictory factors involved are not resolved in a simple, consistent way. Optimality Theory permits the conflicts to be formally adjudicated, and allows the conflicts to be resolved in different ways in different languages. We will present what to our knowledge is the first formal theory of the cross-linguistics typology of discourse anaphora systems.

1 Introduction to the framework

This work builds very directly on much previous research on discourse anaphora and bidirectional Optimality Theory (Beaver 2002, Blutner 2000, de Hoop and de Swart 2000, Hendriks and de Hoop 2001, Smolensky 1996a) and on early pragmatic work which discusses the crucial role of the speaker’s notion of both interpretation and expression (e.g., Grice’s *Logic and Conversation*, 1975). These connections are explicitly identified and discussed in later sections of the paper. But first, we begin with an intuitive introduction of how discourse is conceived in the framework we propose, called ROT for *Recoverability Optimality Theory*. The concepts introduced in this introductory section will be formally defined in the next section.

Consider the following discourse fragments. Each begins with (1a) and (1b), but sentence (1c) varies from **t1** to **t3**.¹

(1) a. Alice gave Betsy a bloody nose.

- b. She_A reminded her_B that her right hook was devastating.
- c. **t1**. She_A hurt her_B.
t2. Betsy hit her_A back.
t3. Alice hurt her_B.
- d. She_A felt awful.

The questions of interest to us are:

- Q1 What principles determine the referent of each pronoun?
- Q2 What are the intermediate and final states of the discourse?
- Q3 Given a communicative intention, why does a speaker choose to reduce some NPs (to overt pronouns or null elements) but not others?

We take the native speaker intuitions concerning the pronoun referents — our data — to be those indicated by the subscripts in (1c), with A/B obviously connoting Alice/Betsy. Consider the case of **t1**: why is the interpretation not “She_B hurt her_A”? In plain English, ROTs answer is this: In the state of the discourse when **t1** is uttered, Alice is more salient than Betsy. The Subject is the most prominent argument and so gets its referent from the most salient discourse entity.

In ROT terminology, the state of the discourse when **t1** is uttered is encoded by the *Salience List* $SaL_{i-1} = [A(\text{lice}) B(\text{etsy})]$. In this simple declarative clause, the Subject occupies the first argument position of the predicate in the *Logical Form* $LF_i = \text{hurt}(A, B) = V(A, B)$.² The alignment of the prominent Subject with the most salient entity favors the pairing of the first (left-most) element of SaL_{i-1} with the first argument of LF_i . This preference is asserted by the ROT constraint $\text{LINEARITY}(LF_i, SaL_{i-1})$; when formalized in 2, it will be clear that this constraint is indeed the standard LINEARITY constraint of OT’s Correspondence Theory McCarthy and Prince (1995), applied to a new kind of correspondence relation.

²We assume throughout that LF preserves information about the grammatical roles of its arguments, and that the overt NPs in our sentences are case-marked so there is no ambiguity about their argument positions, i.e., thematic roles.

* We would like to thank the anonymous reviewers for their valuable insights and helpful comments.

¹Our intention in designing examples is that semantics/pragmatics not bias interpretations, leaving interpretations free to be governed by discourse syntax. For some readers, achieving this may require altering the particular verbs or arguments in the examples we give.

With **t2** in place of **t1**, we see that now it is the Direct Object that refers to A; now $LF_i = V(B, A)$ even though SaL_{i-1} must be the same as before: [A B]. When the Subject is a pronoun “p” = “she” as in **t1**, the most salient discourse entity binds it; when the Subject is the Full NP “B” as in **t2**, the overt information overrides the salience-preference of A from the discourse, and the Subject is interpreted as B. This leaves the Object pronoun “p” needing a referent, which it can get from the most salient discourse entity, A.

ROT conceives this as follows. All features in the interpretation $LF = V(x, y)$ must be licensed, or checked, by a matching feature which resides either in the overt expression uttered – the *Phonetic Form*, PF_i – or in the previously established discourse entities comprising SaL_{i-1} . As we have seen, constraints favor checking features with PF_i if possible; resort to SaL occurs only when PF is insufficient. (Sensible enough given that the overt expression is rather more unambiguously constraining.) So for **t2**, in $V(x, y)$ it is preferable to check the features of x with the Subject of PF_i , which is possible iff $x = B$. A feature that is checked with PF_i will by convention be denoted by a Greek letter, one checked with SaL_{i-1} , a Latin letter. The preference for PF-checking is expressed in ROT simply as $S\Phi = \text{Don't check a feature with SaL}$.

The distinction between features available in PF and those unavailable is formalized in ROT as follows. Nominal features will be divided into two groups: those realized in an overt pronoun “p” (e.g., [gender = F]), p*-features, and those realized in a Full NP but not in a “p” (e.g., [name = Betsy]), f*-features. We denote the former features by subscripts and the latter by superscripts; thus ‘x = B’ means something like $x = X_{[gender=F]}^{[name=Betsy]}$. So if B is expressed with a pronoun, the subscript features are licensed by PF and only the superscript feature need be licensed by SaL: this violates a constraint $S\Phi^k$. If B were expressed with a null element, $S\Phi_k$ would again be violated and in addition another constraint $S\Phi_k$ would be violated: now even the subscript features must be licensed by a discourse entity. When B is expressed by a full NP, none of the $S\Phi$ constraints are violated since all features are licensed by PF.

Consider now (1d). Why is the sole pronoun “p” = “she” interpreted as $V(A)$ and not $V(B)$? Recall that in the interpretation $V(x)$, x must have its superscript features licensed in SaL since the pronoun does not provide them in PF. The primary force of the term ‘salience’ is precisely that when features need to be licensed by the discourse, the features of the most salient entity in SaL are preferred; or equivalently, the features of less salient entities are more costly. In standard OT fashion this will be

formulated as the universal fixed ranking: $[S\Phi_2 = \text{Don't license features by the second element of SaL}] \gg [S\Phi_1 = \text{Don't license features by the first element of SaL}]$.

Turning to Q2, we ask, why does the speaker not give the hearer a break, sparing her $S\Phi$ violations altogether, by always using full NPs? In ROT, the constraint opposing full overt feature expression is simply the all-encompassing economy constraint of OT: $*STRUC = \text{There is no structure}$. The ranking of $*STRUC$ will determine how insistent the speaker is on economizing overt expression by reduction from full NPs to overt pronouns “p” or to null pronouns “ \emptyset ”. In standard OT fashion, the three-way distinction in quantity of structure – Full NP > “p” > “ \emptyset ” – is encoded in $*STRUC$ via the universal hierarchy $*FULL-NP \gg *PRON(\text{or } *STRUC^f \gg *STRUC_{p-\emptyset})$: a full NP violates both, “p” only $*STRUC^f$, and “ \emptyset ” neither). We will compactly notate the PF Betsy hit her back as “BVp” – quotes included; the order is Subj V Obj, “A/B” denote the full NPs (Alice and Betsy), and “p” denotes an overt pronoun. (Null elements will be denoted “ \emptyset ”).

Given $*STRUCTURE$'s pervasive pressure to reduce, why are all NPs not totally reduced? In ROT, the answer is: for the Speaker the intended LF must be recoverable from the uttered PF, and if PF is overly reduced, it will be interpreted with the wrong LF. This is formalized via bidirectional optimization; to a first approximation, it works like this. The expressive optimization Exp is the standard OT optimization in which different PF expressions compete to express a given LF interpretation. In interpretive optimization, Int , different LFs compete as interpretations of a given PF. The Speaker uses Exp to select the optimal PF to express their communicative intention, but all candidate expressions must pass a “pre-screening” Int that eliminates all PFs that will fail to yield the intended interpretation.³ For Exp , what varies across candidates is PF_i ; what is fixed, given — called the index of the optimization — is the discourse state after the previous utterance, SaL_{i-1} , and an intended interpretation LF_i , and an intended new SaL_i : as we see next, different expressions of the same LF will leave the discourse in different states; the optimal expression depends in part on the intended new SaL. For Int , what varies is the new interpretation LF_i and the new discourse state SaL_i ; the given index is the discourse context SaL_{i-1} and the expression PF_i .

Finally, we take up Q3 and consider **t3**. What is the final state of the discourse? As a diagnostic, we can imagine adding a sentence after **t3**, e.g., (1d). Who does “p” refer to now? B. This tells us that after **t3** the salience list is $SaL = [B A]$, because from

³Pre-screening in a logical sense only: we make no claims about the time course of processing.

our discussion of (1d) above we know that a single pronoun must be interpreted as the most salient discourse entity (matching the pronoun’s features). After **t3** B is most salient, even though the Subject of **t3** is A and even though A was more salient before **t3**. This is the power of pronominalization to shift salience. Because the pronoun must take its superscript features from SaL, it is optimal to place B at the head of SaL, where checking features is least costly. Note that the LF of both **t1** and **t3** is V(A, B), but they affect saliency differently: the resulting SaLs are [A B] and [B A], respectively.

Formalizing the intuitive accounts of this section is a major challenge. The next section systematically lays out the proposed solution, the principles of ROT.

2 The formal architecture of ROT

Before presenting our analysis of English and our typological results, a few of the concepts laid out in 1 require some elaboration. After the necessary discussions, the formalization of the system of constraints used in ROT is presented.

2.1 Recoverability and Bidirectionality

ROT is a theory of grammar. It defines the set of utterances available to discourse participants and the set of discourse conditions associated with those utterances.⁴ ROT employs standard OT. In standard OT Prince and Smolensky (1993), Exp models production (performed by a speaker). The architecture of ROT is also derived by considering the point of view of the speaker. The *input* of Exp is an underlying form, and the *output* is the surface phonetic form (Smolensky 1996b). 161996bSmolensky) proposes the optimization be run in the opposite direction (Int), holding the surface form fixed as the index so that the set of candidates varies in their underlying form. This was intended as a model of comprehension (performed by a hearer). Importantly, the set of constraints is presumed to be identical in both directions of optimization.⁵

As described in 1, ROT uses the two directions of optimization to model a speaker’s desire to ensure that his intended interpretation is recoverable from the surface form produced. Because ROT takes the perspective of the speaker, Int is essentially the speaker’s model of the listener’s interpretive process.⁶ As such, the speaker has access to the results and

⁴Whether ROT is a viable model of language production is left as an open question for future work.

⁵Recent work in OT-Semantics Hendriks and de Hoop (2001), Zeevat (2000), inter alia, has taken a contrary position on this.

⁶The authors subscribe to the (possibly controversial) view that a speaker’s utterances are influenced by her evaluation of whether or not she will be understood to have said what she meant. However, we do not claim that the speaker’s model

may examine the most harmonic candidate of Int, evaluated over the candidate phonological forms under consideration in Exp, to ensure that the most optimal form (evaluated in Exp), whose meaning is recoverable (evaluated in Int) is chosen to convey the intended meaning.

Following Wilson (2002), we employ a version of bidirectional OT that crucially relies on the interaction between directions of optimization. This interaction is motivated the desire of the speaker to ensure that the intended utterance is recoverable from the surface form produced.⁷ The architecture can be formally restated as:

- (2) In producing an utterance for an intended interpretation $I = \langle LF_i, SaL_i \rangle$, a speaker will use the form $F = PF_i$ of the most most harmonic candidate $o = \langle I, F \rangle$, evaluated by $\text{Exp}(I)$, such that there is no candidate $o' = \langle I', F' \rangle \succ o$ (as evaluated by $\text{Int}(F)$).

This formulation of bidirectional OT falls in between the “strong” and “weak” versions considered by Blutner (2000). In Blutner’s terms, “strong” bidirection admits pairs $\langle I, F \rangle$ iff $F \in \text{Exp}(I)$ (strong Q-principle) and $I \in \text{Int}(F)$ (strong I-principle). “Weak” bidirection uses recursive versions of these principles:

- (3) Q-principle: $\langle I, F \rangle$ satisfies Q iff $\langle I, F \rangle$ satisfies I and $F \in \text{Exp}(I)$.
I-principle: $\langle I, F \rangle$ satisfies I iff $\langle I, F \rangle$ satisfies Q and $I \in \text{Int}(F)$.

In order for “strong” and “weak” versions to be distinct requires an additional stipulation of the architecture: there must be a one-to-one correspondence between form and meaning. This causes an interpretation I_1 not to compete in $\text{Int}(F_2)$ if there is a pair $\langle I_1, F_1 \rangle \succ \langle I_1, F_2 \rangle$, which allows the less harmonic pair $\langle I_2, F_2 \rangle$ will to be admitted into the language. This generates the phenomenon that “marked forms have marked meanings,” (often referred to as *partial blocking*) The most important difference between our conception of bidirection and Blutner’s is that we do not stipulate this one-to-one correspondence between form and meaning as a part of the architecture. ROT uses the weak Q-principle, but the strong I-principle, which allows for the possibility of

of the hearer’s interpretation of a phonological form (which is the version of Int employed by ROT) is necessarily consistent with the hearer’s model of the same.

⁷Note that in the following definition, we omit SaL_{i-1} from the notation, understanding that as the prior context for a given utterance, SaL_{i-1} is fixed and is present in the indices of both directions of optimization for that utterance; note additionally that the shorthand $\langle a, b \rangle \in \text{Int}(b)$ means that there is no pair $\langle a', b \rangle$ s.t. $\langle a', b \rangle$ is more harmonic than $\langle a, b \rangle$, as evaluated by $\text{Int}(b)$ i.e. that $\langle a, b \rangle$ does not lose the interpretive optimization over b .

ambiguous interpretations. Although adding the requirement that there be a one to one correspondance between form and meaning gives an account of partial blocking Blutner (2000), the same stipulation rules out neutralization phenomena, in which different underlying forms *do* have identical surface forms (in production).⁸ It seems to us that the subtleties captured by imposing a one to one correspondance could be left to the mechanics of the constraints, and that an architecture should not rule out by definition any linguistic phenomena. Although at present we do not provide an account of partial blocking, our architecture does not seem to *a priori* rule this out.

2.2 Features, Topic, & Salience: from Centering Theory to ROT

Much of the inspiration for ROT derives from Centering Theory (CT: Grosz, Joshi, and Weinstein (1995)). In CT, the discourse context consists of a list of discourse entities (Centers), whose default ordering may vary from language to language. CT analyzes a discourse in terms of a sequence of local “discourse transitions” from one utterance to the next. The type of discourse transition between any two utterances is defined in terms of the changes in the values of (hidden) variables (backward- and forward-looking Centers), and organized in terms of the “local coherence” of discourse segments. These variables are themselves constructed concepts within CT, and their values are not always readily apparent. Given the fact that CT presupposes a good deal of syntactic and semantic analysis, we hypothesized that a careful, pragmatic treatment of the concept of Recoverability at the interface between syntax, semantics, and discourse would derive the same core set of results as Centering Theory.

Beaver (2002)’s important work on redefining CT in an OT framework (COT), replaces the Centers of CT with a system of constraints built around the concept of “Topic,” an often used but still controversial notion.⁹ We agree with Beaver’s intuition that “topichood” is not always easy to unambiguously define. He suggests an OT analysis that selects a topic based on the more well founded concept of “salience.” ROT takes another step in this direction, by dispensing of the notion of “topic” altogether, and allowing the whole system emerge from the interactions between constraints that do not depend on special-status elements. The motivation for our set of $S\Phi$ constraints is essentially the same as Beaver’s salience-based analysis of topic. In future work, we hope to augment ROT with a richer set

of features (such as empathy, -wa marking, cue validity, accessibility, etc.) and corresponding sets of constraints that belong to the $S\Phi$ family.

The specific types of features ROT presently employs are mostly left unspecified, given the restriction of the present analysis to this limited domain. At present, they are simply meant to correspond to features that may be present or absent from the particular referring expressions under consideration (such as gender, number, or name). Their ability to restrict the set of possible referents of the anaphoric expression is currently the relevant factor.

2.3 Features, Interpretation & Expression

As discussed in 1, in order for an utterance U_i to be interpreted, the entities in the argument positions of LF_i must have each of their features licenced by corresponding features in SaL_{i-1} or the overt form PF_i . This deserves some elaboration. SaL_{i-1} is meant to correspond to the speaker’s internal model of the *common ground*. We assume (for now) that all of the entities in SaL_{i-1} appear there with all of their features. As such, in Int , it will always be possible to license any feature of an anaphoric entity that does not appear in PF_i , albeit incurring a cost via $S\Phi$ constraint violations. This makes sense in the (speaker’s model of a hearer’s) interpretive process, as it seems natural to require full feature set specification (or at least enough to uniquely identify the referents) for interpretation,¹⁰ and for the speaker, such features are presumably available.

ROT also requires each feature of LF_i to be licenced in Exp , although this has a slightly different meaning. In Exp , which corresponds to production, the licensing requirement is more naturally thought of as checking each feature to ensure that the entire LF_i is actually expressed. Accordingly, the licensing of features in Exp takes place between LF_i , PF_i , and SaL_i . The identity of elements in SaL_i and their relative ordering is specified in the index of Exp , but since features are only required to be there there if they are not in PF_i (which varies among the set of candidates), SaL_i is not assumed to have features where they are not necessary. Rather, the entities in SaL_i receive the rest of their features as a consequence of becoming a part of the speaker’s model of the *common ground* for the next utterance U_{i+1} .

2.4 Constraints

The constraints in ROT come from three families, and are formally defined below. As mentioned, all constraints are assumed to be present in every optimization. As in OT, each possible re-ranking of constraints (i.e., those that do not violate universal subhierarchies) generates a potential grammar of a natural language.

¹⁰Leaving aside for the moment the complexities of under-specification in semantic or discourse theories.

⁸For example, German requires devoicing of syllable-final consonants, which leads to the same relationship, e.g., /rat/ → /rat/ and /rad/ → /rat/

⁹See Jennifer Arnold’s thesis (Arnold 1998) for a comprehensive review of different notions of Topic in the literature, and the relationships between Topic, Focus and salience.

Recoverability

In ROT, the $S\Phi$ constraints are often called “recoverability” constraints. $S\Phi$ constraints penalize the use of less salient positions for feature licensing. For the present feature set, this yields:¹¹

- (4) $S\Phi^i$: do not license a p^* -feature with one at position i in SaL .
 $S\Phi_i$: do not license a f_* -feature with one at position i in SaL .

with the universal rankings: $S\Phi^m \gg S\Phi^n$ iff $m > n$, and $S\Phi_m \gg S\Phi_n$ iff $m > n$.

Linearity

We use two constraints from the LINEARITY family of Faithfulness constraints (McCarthy and Prince 1995) that penalize candidates when the order of features (according to the hierarchy of grammatical role) that are in LF_i (but not PF_i) does not align linearly with the order of elements in SaL_{i-1} or SaL_i .

- (5) $LIN\Phi_{i-1}$: For a pair of features x, y in both Sf and SaL_{i-1} , x precedes y in SaL_{i-1} iff x precedes y in Sf .
 $LIN\Phi_i$: For a pair of features x, y in both Sf and SaL_i , x precedes y in SaL_i iff x precedes y in Sf .

In ROT, these constraints serve a double role: they introduce the relationship between salience and grammatical role;¹² and they ensure (indirectly and fallably) the coherence of successive utterances (their ultimate preference is for the orders of entities in SaL_{i-1} and SaL_i to align with each other, (and with obliqueness)).

Economy

Finally, we also use two economy constraints from the *STRUC constraint family that penalize phonological material in the candidate:

- (6) *FULL: No Full-NPs
 *PRON: No pronouns

There is a universal ranking among these two constraints, such that *FULL \gg *PRON.

3 The Mechanics

In this section, we illustrate the ROT account of discourse anaphora using examples from English. We assume throughout that a speaker has a fixed SaL_{i-1} , intended interpretation LF_i , and target SaL_i for any U_i . ROT defines four transitions over these dimensions:

¹¹More generally, for each feature ϕ in the system, there is a corresponding set of $S\Phi\phi, k$ constraints (where k indexes the position in SaL used to license feature ϕ), with the universal ranking $S\Phi\phi, m \gg S\Phi\phi, n$ iff $m > n$.

¹²*Ceteris paribus*, the central cross-linguistic factor that determines the ordering of the Cf lists of CT (Grosz et al. 1995); also similar to the effects of the SALIENTARG constraint of COT Beaver (2002).

- t1**: $SaL_{i-1} = SaL_i$ AND LF_i is aligned with SaL_{i-1}
t2: $SaL_{i-1} = SaL_i$ AND LF_i is NOT aligned with SaL_{i-1}
t3: $SaL_{i-1} \neq_{SaL} [i]$ AND LF_i is aligned with SaL_{i-1}
t4: $SaL_{i-1} \neq SaL_i$ AND LF_i is NOT aligned with SaL_{i-1}

The following discourse segment from English illustrates these transitions:

- (7) a. Alice gave Betsy a bloody nose.
 b. She_A reminded her_B that her right hook was devastating.
 c. **t1**. She_A hurt her_B.
 t2. Betsy hurt her_A.
 t3. Alice hurt her_B.
 t4. Betsy hurt her_A.

ROT represents the transitive sentences in (7c) as the four-tuples in (3):¹³

- (8) $\langle SaL_{i-1}, SaL_i, LF_i, "PF_i" \rangle$
 a. **t1**. $\langle [A, B], [A, B], V(A, B), "pVp" \rangle$
 b. **t2**. $\langle [A, B], [A, B], V(B, A), "BVp" \rangle$
 c. **t3**. $\langle [A, B], [B, A], V(A, B), "AVp" \rangle$
 d. **t4**. $\langle [A, B], [B, A], V(B, A), "BVp" \rangle$

To illustrate properties of ROT discussed above, we will demonstrate how the PFs for **t1** and **t2** are selected for English-type languages.

3.1 Int

As mentioned, ROT contains a formal mechanism to ensure that candidate expressions yield the intended interpretation: the evaluation in **Int**. Tableau 1 shows **Int** for **t1**, with an index of “she hurt her” (or “pvp”) and SaL_{i-1} of $[A, B]$. The constraint ranking in Tableau 1 is the ranking that generates English. The quivered arrow symbol points to the most harmonic candidates in **Int**, representing the optimal interpretation.

In Tableau 1, the most harmonic candidate is (a), corresponding to the four-tuple **t1**. Candidates (b-d) each incur fatal violations of LINEARITY constraints: for candidates (c) and (d) the order of the features in LF that are not accessible from PF differs from the order in SaL_{i-1} ; for candidates (b) and (c) they differ from the order in SaL_i . These candidates are *harmonically bound* by candidate (a), and will not be optimal under any constraint ranking. According to principles of OT that are employed in ROT,

¹³ROT has two transitions (**t2** and **t4**) that have the same PF. This occurs as there can be multiple candidates in **Int** for which no candidate that is more harmonic in the evaluation. **Int** evaluates the $S\Phi$ constraints over SaL_{i-1} , so (in some cases) two candidates that differ only in SaL_i will be equally harmonic.

this result will hold cross-linguistically; that is, ROT predicts that no language uses two pronouns to represent a change from SaL_{i-1} to SaL_i , or when the LF does not align linearly with SaL_{i-1} . The explanation afforded by ROT is that these interpretations are not recoverable from the PF.

SaL_{i-1} : [A',B']	SaL_i : [A',B']	LF: V(A',B')	PF: "p _n Vp _n "	"She _n hurt her _n "	LIN	LIN Φ	SF ₂	SF ₁	*FU	*PR	SF ²	SF ¹
a. LF: V(A _n ¹ ,B _n ²)	SaL_i : [A ¹ ,B ²]	hurt (Alice _n ¹ , Betsy _n ²)							**	*	*	*
b. LF: V(A _n ¹ ,B _n ²)	SaL_i : [B ² ,A ¹]	hurt (Alice _n ¹ , Betsy _n ²)				*!			**	*	*	*
c. LF: V(B _n ² ,A _n ¹)	SaL_i : [A ¹ ,B ²]	hurt (Betsy _n ² , Alice _n ¹)			*!	*			**	*	*	*
d. LF: V(B _n ² ,A _n ¹)	SaL_i : [B ² ,A ¹]	hurt (Betsy _n ² , Alice _n ¹)			*!				**	*	*	*

Tableau 1: $\text{Int}(SaL_{i-1}=[A, B]; \text{PF}=\text{"pVp"})$

3.2 Exp

Recall that **Exp** evaluates PFs (much like standard OT), holding everything else in the candidate fixed. Tableau 2 shows **Exp** with an index SaL_{i-1} and SaL_i of [A,B], and LF_i of $V(A,B)$. Although LF remains constant for each candidate PF, the features in LF are licensed differently; to facilitate the reader, the licensing for features in LF_i is specified within each candidate in Tableau 2, and the pattern of violations can be read off from the numerical sub- and superscripts of the candidate LFs. For example, candidate (a) contains full NPs for both arguments (and no numerical indices), so no $S\Phi$ constraints are violated, whereas both f^* - and p_* -features of candidate (e) are licensed by SaL_i (as seen in the indices), so this candidate violates each $S\Phi$ constraint.¹⁴ Candidate (b) (the four-tuple from **t1**) is the optimal output in **Exp** under the constraint ranking of English. The crucial ranking ($S\Phi^f \gg \text{PRON} \gg S\Phi_p$) reflects the fact that, in English, pronominalizing is preferred to full NPs, but having p^* -features licensed by an element in SaL_i (as with θ -forms) is worse than using a pronoun.

SaL_{i-1} : [A',B']	SaL_i : [A,B]	LF: V(A,B)	PF: "p _n Vp _n "	"Hurt (Alice, Betsy)"	SF ₂	SF ₁	*FU	*PR	SF ²	SF ¹
a. PF: "A _n ¹ Vp _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A,B]						*!			
b. PF: "p _n Vp _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B']							**	*	*
c. PF: "p _n Vp _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B]						*!	*	*	*
d. PF: "A _n ¹ Vp _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A,B]						*!	*	*	*
e. PF: "∅V∅∅"	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B ²]				*!	*			*	*
f. PF: "∅VB _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B]					*!	*		*	*
g. PF: "A _n ¹ V∅∅"	LF: V(A _n ¹ ,B _n ²); SaL_i : [A,B ²]				*!		*		*	*
h. PF: "∅Vp _n ² "	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B ²]				*!			*	*	*
i. PF: "p _n V∅∅"	LF: V(A _n ¹ ,B _n ²); SaL_i : [A',B ²]					*!		*	*	*

Tableau 2: $\text{Exp}(SaL_{i-1}=[A,B]; SaL_i=[A,B]; LF=V(A,B))$

3.3 The Interaction of Int and Exp

As discussed earlier, English speakers who want to change the LF and reorder the SaLs use utterances that conform to the four-tuple identified in **t2**. Tableau 3 shows **Exp** for this situation.

¹⁴As each candidate violates $\text{LIN}\Phi_{i-1}$ and satisfies $\text{LIN}\Phi_i$, these constraints do not figure in the outcome of this optimization, and are omitted. The gray candidates in Tableau 2 are those that do not yield the intended interpretation (see Appendix for the full description of Int).

SaL_{i-1} : [S',B']	SaL_i : [B,S]	LF: BSV	SF ₂	SF ₁	*FU	*PR	SF ²	SF ¹
a. V(B _n ² ,S _n ¹); "p _n Vp _n ² "	SaL_i : [B,S]				**!		*	*
b. V(B _n ² ,S _n ¹); "p _n Vp _n ² "	SaL_i : [B',S']					**	*	*
c. V(B _n ² ,S _n ¹); "p _n Vp _n ² "	SaL_i : [B',S]				*	*	*	*
d. V(B _n ² ,S _n ¹); "B _n ² Vp _n ² "	SaL_i : [B,S']				*	*	*	*
e. V(B _n ² ,S _n ¹); "∅V∅∅"	SaL_i : [B ² ,S ₁ ¹]		*	*			*	*
f. V(B _n ² ,S _n ¹); "∅VS _n ¹ "	SaL_i : [B ² ,S']		*	*		*	*	*
g. V(B _n ² ,S _n ¹); "B _n ² V∅∅"	SaL_i : [B,S ₁ ¹]			*!	*	*	*	*
h. V(B _n ² ,S _n ¹); "∅Vp _n ² "	SaL_i : [B ² ,S']		*	*		*	*	*
i. V(B _n ² ,S _n ¹); "p _n V∅∅"	SaL_i : [B',S ₁ ¹]		*	*		*	*	*

Tableau 3: $\text{Exp}(SaL_{i-1}=[A,B]; SaL_i=[B,A]; LF=V(B,A))$

Of the candidates in Tableau 3 that are not neutralized in **Int**, the constraint alignment of English favors candidate (d), the four-tuple **t2**. The need to license p-features in PF rules out candidate (g), and candidate (a) incurs two violations of *FULL-NP, the second being fatal.

Tableaux 1 and 3 demonstrate the import of the interaction between **Int** and **Exp**. As stated, all candidates that compete in **Exp** must yield the target interpretation. Although candidate (b) would be the most harmonic candidate in Tableau 3 with this constraint ranking, it does not compete in **Exp** as it was ruled out in Tableau 1 (candidate (d)). This demonstrates the "bidirectionality" of the ROT architecture.

In this section, we have used a basic analysis of the use of English anaphora to demonstrate how the ROT architecture generates a grammar. Crucially, the interaction of **Int** with **Exp** under the same constraint ranking rules out certain candidates that do not yield the target interpretation. In the next section, we will see how properties of constraint interaction inherent in OT provide a framework for thinking about the cross-linguistic variation in the use of anaphora.

4 Typology in ROT and other considerations

Inherent in any OT analysis are predictions about cross-linguistic typology. Constraints in OT can be re-ranked, and each possible ranking is a *potential* natural language. To assess the typological predictions of ROT, we examine the results of the full set of **Int** and **Exp** for each possible constraint ranking.¹⁵ ROT generates several well-attested language types, listed in Table 3.1. L1 requires Full NPs for any of the four transitions. We refer to this language as children's storybook English. Note that there is no ambiguity in this language, since any PF of the utterance will crucially depend only on LF. English, as discussed in the previous section, fits type L2, although it may also fit type L3. Japanese and Turkish seem to fit the pattern outlined in L4, in which zero pronouns are used for salient discourse referents (Walker, Iida, and Cote 1994, Kameyama 1998, Turan 1995). If we consider pro to be a zero, Italian

¹⁵This diverges from other prominent treatments of semantics using bidirectional OT (e.g., Beaver 2000, Hendriks and de Hoop 2000, Blutner 2000).

and Greek fit the description of L7 (and possibly L6; Di Eugenio (1990, 1998), Dimitriadis (1996)).¹⁶

Table 1: Typology

L	t1	t2	t3	t4	Exemplar
L1	"AVB"	"BVA"	"AVB"	"BVA"	Storybook
L2	"pVp"	"BVp"	"AVp"	"BVp"	English(1)
L3	"pVB"	"BVp"	"AVp"	"BVp"	English(2)
L4	" \emptyset V \emptyset "	"BV \emptyset "	" \emptyset VB"	"BV \emptyset "	Japanese
L5	" \emptyset VB"	"BVp"	"AVp"	"BVp"	
L6	" \emptyset Vp"	"BV \emptyset "	"AVp"	"BV \emptyset "	
L7	" \emptyset Vp"	"BVp"	"AVp"	"BVp"	Italian

Each of these languages represents a particular ranking of constraints such that L1's output is realized by a ranking of all $S\Phi$ constraints over *STRUC constraints ($S\Phi^{1,2} \gg^* \text{STRUC}$) and L4 and L7 are generated by $S\Phi^2 \gg^* \text{FULL} \gg^* \text{PRON} \gg^* S\Phi^1$ and *FULL $\gg^* \text{PRON} \gg^* S\Phi^2$ respectively.

The predictions of ROT are necessarily limited by the distinctions among features and expressions (e.g., Full NP, pronoun, null). As mentioned in section 2, ROT is designed to be augmented. For example, future work may include expressions "in between" pronoun and zero, such as subject agreement on a verb, or clitics. Alternatively, identifying further classes of features would allow us to use finer-grained $S\Phi$ constraints. Changes such as these are necessary to broaden the predicted typology, allowing ROT to become more precise in its cross-linguistic account of anaphora. At present, ROT seems promising as a framework for the study of linguistics at the interfaces of several of its subfields.

References

Arnold, Jennifer. 1998. Reference form and discourse patterns. Doctoral Dissertation, Stanford University.

Beaver, David. 2002. The optimization of discourse anaphora. *Linguistics and Philosophy* To appear.

Blutner, Reinhard. 2000. Some aspects of optimality in natural language interpretation. In de Hoop and de Swart (2000), 1–21.

Clark, H. H., and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, ed. L. B. Resnick, J. M. Levine, and S. D. Teasley, 127–149. APA Books.

Di Eugenio, Barbara. 1990. Centering theory and the Italian pronominal system. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, 270–275.

Di Eugenio, Barbara. 1998. Centering in Italian. In Walker, Joshi, and Prince (1998).

Dimitriadis, Alexis. 1996. When pro-drop languages

don't: Overt pronominal subjects and pragmatic inference. In *CLS 32: The Main Session*, ed. Lise M. Dobrin, Kora Singer, and Lisa McNair, 33–47. Chicago Linguistic Society.

Grice, H. P. 1975. Logic and conversation. In *Syntax and Semantics*, ed. P. Cole and J. L. Morgan, 41–58. Academic Press.

Grosz, Barbara A., Aravind Joshi, and Scott Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* 21:203–226.

Hendriks, P., and Helen de Hoop. 2001. Optimality theoretic semantics. *Linguistics and Philosophy* 21:1–32.

de Hoop, Helen, and Henriette de Swart, ed. 2000. *Papers on Optimality Theoretic Semantics*. Utrecht Institute of Linguistics OTS.

Kameyama, Megumi. 1998. Intrasentential centering: A case study. In Walker et al. (1998).

McCarthy, John J., and Alan S. Prince. 1995. Faithfulness and reduplicative identity. Ms., U. Mass. and Rutgers, July 1995.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, New Brunswick, NJ and University of Colorado at Boulder.

Smolensky, Paul. 1996a. Generalizing optimization in ot: A competence theory of grammar 'use'. Paper presented at the Stanford Workshop on Optimality Theory, Stanford University.

Smolensky, Paul. 1996b. On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27:720–731.

Turan, Ümit Deniz. 1995. Null vs. overt subjects in Turkish discourse: A centering analysis. Doctoral Dissertation, University of Pennsylvania.

Walker, Marilyn, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 21:193–232.

Walker, Marilyn, Aravind Joshi, and Ellen Prince, ed. 1998. *Centering theory in discourse*. New York: Oxford University Press.

Wilson, Colin. 2002. Bidirectional optimization and the theory of anaphora. In *Optimality-theoretic Syntax*, ed. Géraldine Legendre, Jane Grimshaw, and Sten Vikner. Cambridge: MIT press.

Zeevat, Hank. 2000. The asymmetry of optimality theoretic syntax and semantics. *Journal of Semantics* 17:243–262.

¹⁶Miltsakaki (2001) notes that there are three possible reduced realizations for entities in Greek (as in many other languages) pro, weak pronouns, and strong pronouns, our typology does not yet handle this degree of variation. Nonetheless, we hope we have made clear that such variation is easily dealt with in ROT.

5 Appendix

I1	SaL _{i-1} = [A ¹ , B ²]; "AVB"	Li (i-1)	LiN (i)	*F U	*P R	*X ²	*X ¹
⇒	a. [A, B]; V(A _α ^μ , B _β ^ν)			**			
⇒	b. [B, A]; V(A _α ^μ , B _β ^ν)			**			

I2	SaL _{i-1} [A ¹ , B ²]; "pVB"	LN (i-1)	LN (i)	*F U	*P R	*X ²	*X ¹
⇒	a. [A ¹ , B]; V(A _α ¹ , B _β ^ν)			*	*		*
⇒	b. [B, A ¹]; V(A _α ¹ , B _β ^ν)			*	*		*
	c. [B]; V(B _α ² , B _β ^ν)	*		*	*	*	

I3	SaL _(i-1) = [A ¹ , B ²]; "AVp"	LIN (i-1)	LIN (i)	*F U	*P R	*X ²	*X ¹
⇒	a. [A, B ²]; V(A _α ^μ , B _β ²)			*	*	*	
⇒	b. [B ² , A]; V(A _α ^μ , B _β ²)			*	*	*	
	c. [A]; V(A _α ^μ , A _β ¹)	*		*	*	*	*

I4	SaL _(i-1) = [A ¹ , B ²]; "BVp"	LIN (i-1)	LIN (i)	*F U	*P R	*X ²	*X ¹
⇒	a. [B, A ¹]; V(B _α ^μ , A _β ¹)	*		*	*		*
⇒	b. [A ¹ , B]; V(B _α ^μ , A _β ¹)	*		*	*		*
	c. [B]; V(B _α ^μ , B _β ²)	*		*	*	*	

I5	SaL _(i-1) = [A ¹ , B ²]; "pAV"	LIN (i-1)	LIN (i)	*F U	*P R	*X ²	*X ¹
	a. [B ² , A]; V(B _α ² , A _β ^ν)	*		*	*	*	
	b. [A, B ²]; V(B _α ² , A _β ^ν)	*		*	*	*	
⇒	c. [A]; V(A _α ¹ , A _β ^ν)	*		*	*		*

I6	SaL _(i-1) = [A ¹ , B ²]; "BVA"	LIN (i-1)	LIN (i)	*F U	*P R	*X ²	*X ¹
⇒	a. [B, A]; V(B _α ^μ , A _β ^ν)	*		*	*		
⇒	b. [A, B]; V(B _α ^μ , A _β ^ν)	*		*	*		

I7	SaL _(i-1) = [A, B]; "pVØ"	LIN i-1	LIN (i)	*P R	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(A ¹ , B ₂ ²)			*	*		*	*
	b. [B, A]; V(A ¹ , B ₂ ²)		*	*	*		*	*
⇒	c. [B, A]; V(B ² , A ₁ ¹)	*		*	*	*	*	*
	d. [A, B]; V(B ² , A ₁ ¹)	*	*	*	*	*	*	*

I8	SaL _(i-1) = [A, B]; "AVØ"	LIN i-1	LIN (i)	*F U	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(A, B ₂ ²)			*	*		*	
⇒	b. [B, A]; V(A, B ₂ ²)			*	*		*	
	c. [A]; V(A, A ₁ ¹)	*		*	*	*	*	*

I9	SaL _(i-1) = [A, B]; "ØVp"	LIN i-1	LIN (i)	*P R	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(A ₁ ¹ , B ²)			*	*	*	*	*
	b. [B, A]; V(A ₁ ¹ , B ²)		*	*	*	*	*	*
	c. [B, A]; V(B ₂ ² , A ¹)	*		*	*	*	*	*
	d. [A, B]; V(B ₂ ² , A ¹)	*	*	*	*	*	*	*

I10	SaL _(i-1) = [A, B]; "BVØ"	LIN i-1	LIN (i)	*F U	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(B, A ₁ ¹)	*		*	*	*	*	*
⇒	b. [B, A]; V(B, A ₁ ¹)	*		*	*	*	*	*
	c. [B]; V(B, B ₂ ²)	*		*	*	*	*	*

I11	SaL _(i-1) = [A, B]; "ØVB"	LIN i-1	LIN (i)	*F U	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(A ₁ ¹ , B)			*	*	*	*	*
⇒	b. [B, A]; V(A ₁ ¹ , B)			*	*	*	*	*
	c. [B]; V(B ₂ ² , B)			*	*	*	*	*

I12	SaL _(i-1) = [A, B]; "ØVA"	LIN i-1	LIN (i)	*F U	*X ₂	*X ₁	*X ²	*X ¹
	a. [A, B]; V(B ₂ ² , A)	*		*	*	*	*	*
	b. [B, A]; V(B ₂ ² , A)	*		*	*	*	*	*
⇒	c. [A]; V(A ₁ ¹ , A)			*	*	*	*	*

I13	SaL _(i-1) = [A, B]; "ØVØ"	LIN i-1	LIN (i)	*X ₂	*X ₁	*X ²	*X ¹
⇒	a. [A, B]; V(A ₁ ¹ , B ₂ ²)			*	*	*	*
	b. [B, A]; V(A ₁ ¹ , B ₂ ²)		*	*	*	*	*
	c. [B, A]; V(B ₂ ² , A ₁ ¹)	*		*	*	*	*
	d. [A, B]; V(B ₂ ² , A ₁ ¹)	*	*	*	*	*	*

Exp(t3)	SaL _{i-1} = [A, B]; V(A, B); SaL _i = [B, A]	LIN i-1	LIN (i)	*F U	*P R	*X ₂	*X ₁	*X ²	*X ¹
	a. "AVB"; V(A, B)			**					
	b. "AVp"; V(A, B ₂)			*	*			*	
	c. "ØVB"; V(A ₁ ¹ , B)	*		*			*	*	*