The Optimization of Discourse Anaphora [*]

David I. Beaver
Stanford University

Abstract. In this paper the Centering model of anaphora resolution and discourse coherence (Grosz, Joshi and Weinstein, 1983, 1995) is reformulated in terms of Optimality Theory (OT) (Prince and Smolensky 1993). One version of the reformulated model is proven to be descriptively equivalent to an earlier algorithmic statement of Centering due to Brennan, Friedman and Pollard (1987). However, the new model is stated declaratively, and makes clearer the status of the various constraints used in the theory. In the second part of the paper, the model is extended, demonstrating the advantages of the OT reformulation, and capturing formally ideas originally described by Grosz, Joshi and Weinstein. Three new applications of the extended OT Centering model are described: generation of linguistic forms from meanings, the evaluation and optimization of extended texts, and the interpretation of accented pronouns.

Keywords: semantics, pragmatics, centering, Optimality Theory, anaphora resolution, discourse

1. Overview

In the last twenty years, the fields of formal semantics and pragmatics have seen a great deal of research on the interpretation of anaphora. Of particular note is the development of dynamic approaches to meaning (Kamp and Reyle, 1993; Heim, 1982; Groenendijk and Stokhof, 1991). Yet there has been a curious near absence of work within this tradition on anaphora *resolution*: models have tended to concentrate on absolute semantic constraints on what can be anaphoric to what, rather than to build up detailed pictures of which discourse entities are salient, and hence likely to be referred to, at which times.

David Beaver

There is a separate strong tradition of work on anaphora resolution (Grosz et al., 1983; Sidner, 1983; Grosz and Sidner, 1986; Gundel, 1998). This work is similarly dynamic, in that the core of these models is an account of the impact of an utterance on the information state of conversational participants. With one or two notable exceptions (e.g. Roberts, 1998), work on anaphora resolution has received far more attention in the natural language processing and psycholinguistics communities than from formal semanticists and pragmaticists. The current paper is intended to help bridge the gap between these separate communities.

Centering (Grosz et al., 1983; Grosz et al., 1995) is intended to model discourse coherence, inference by conversational participants, and anaphora resolution. A quite separate line of linguistic research has produced Optimality Theory (OT), a formal framework for reasoning about combinations of linearly ranked linguistic constraints (Prince and Smolensky, 1993). OT has been extraordinarily influential in phonology, has made a significant impact on some areas of syntax (e.g. formal models of typology), and, of particular relevance to the current enterprise, has recently started to make inroads into semantics and pragmatics.[1] I will present a restatement and development of the Centering model in Optimality Theory.

I will begin with presentation of a standard variant of Centering. It is not my purpose to motivate Centering here, for which readers are referred to the original papers. Having stated the archetype theory, I describe the OT reformulation, and then demonstrate the application of the resulting system, COT, to examples. This part of the paper ends by making formally precise the sense in which COT is a reformulation rather than a descriptively new theory.

While the early sections of the paper are conservative with respect to the existing theory of Centering, in the remainder I suggest innovations to the model, and ways in which it might be integrated within an account of discourse processing and conversational inference. Extensions to COT proposed include applications to the evaluation of complete texts, to text generation, and to the interpretation of stressed pronouns. The paper ends with suggestions for further research, and discussion of how the developments in the paper relate to the original goals of Centering theory.

---

[1] Recently OT work on semantics and pragmatics has been blossoming, see Blutner (2000), Blutner and Jäger (1999), Dekker and van Rooy (2000), Hendriks and de Hoop (2001), de Hoop (2000), de Hoop and de Swart (1998), van der Does and de Hoop (1998), Zeevat (1999), Zeevat (2000) and Langendoen (2000).

2. Introduction to Centering Theory

I will now introduce the motivating ideas of Centering theory and describe in detail one particular instantiation of Centering (Brennan et al., 1987). This will provide the departure point for the proposals to be made in later sections of the paper.

2.1. THE ORIGINAL MODEL

The early articles in the Walker et al. (1998) collection provide a good introduction to the theory of Centering. The theory resulted from the fusion of two lines of thought. On the maternal side, it incorporates ideas from Grosz and Sidner's work on anaphora resolution and discourse coherence, work which has appeared in their models of local and global discourse structure (Grosz, 1977; Sidner, 1983; Grosz and Sidner, 1986). The paternal line, from which the framework's name descends, includes work on inference in discourse by Joshi and associates (Joshi and Kuhn, 1979; Joshi and Weinstein, 1981). The first published paper drawing together these lines of thought was that of Grosz et al. (1983), and a more extended presentation of the framework did not appear in print until Grosz et al. (1995). We may sum up the main themes of Centering as follows:

1. The attentional state of language users evolves dynamically through production or comprehension of a discourse, on a sentence by sentence basis. A change of attentional state is known as a *transition*.

2. Attentional state is related to ease of inference: certain inferences associated with salient entities are made more easily than comparable inferences unrelated to salient entities.

3. Transitions in attentional state are classified according to the amount of change involved.

4. Coherence of a discourse is dependent on the transitions made in processing that discourse. Coherent discourses are those in which major attentional changes take place relatively infrequently, and in which such changes are signaled in advance. Incoherence or processing difficulty results when a discourse forces frequent unanticipated refocusing of the hearer's attention.

5. An attentional state determines which entities are under discussion, the *centers* of attention.

6. By considering all ways in which linguistic form may relate to the centers of attention, and trying to maximize coherence of the discourse, we may make predictions about when anaphoric expressions should be used and how anaphora is resolved.

The model presented by Brennan et al. (1987) (BFP) cashes out some of these themes with sufficient precision to produce a predictive model of anaphora resolution.[2] The model provides much of the groundwork for the reformulation I will propose, although it differs in a number of crucial respects. First, BFP, as with all other existing presentations of Centering, is intrinsically procedural. The model I will propose is stated declaratively, although it has a decision procedure. The declarative statement of the theory does not deny the dynamic nature of Centering, but abstracts away from any particular algorithm or heuristics that might be used in an implementation. Second, the models differ in the degree to which linguistic generalizations about anaphora resolution are integrated into a single level of description. Whereas the model I will propose is in this sense highly integrated, in BFP the generalizations of centering are stated at a number of levels, some as absolute constraints on reference, some as transition specifications, and some as preferences between transitions. In the model I will propose transitions are no longer a core part of the theory, although they may still be identified epiphenomenally.

The late Megumi Kameyama has also suggested a system of defeasible constraints that would make transitions epiphenomenal (Kameyama, 1998). Her work, like BFP, is an inspiration behind the current one. I will incorporate her suggestion, but (a) use a different set of constraints and (b) make the model more predictive by stating a constraint ranking. The resulting system allows easy calculation of anaphoric resolution preferences, thus greatly clarifying the work which she began. Despite the relevance of Kameyama's model, it is the more computationally precise BFP model that provides the basis of COT, and the remainder of this section will be taken up with a description of BFP.

2.2. The BFP model

In the Centering model, a sentence provides a mapping from an input information state to an output state. The state represents the sentence's anaphoric potential, and captures the relative salience of various discourse entities. A state comprises a *backward-looking center* (name of

---

[2] It is important to realize that the BFP model differs from the apparent intentions of the original proponents of Centering in a number of ways, as manifested, e.g. the discussion of Centering's "Rule 2" in footnote 4, below. For discussion of the differences between various Centering proposals, see Kehler (1997).

a discourse referent), and a *forward-looking center list* (list of referent names).

The backward center is a link with the previous sentence: it is the most significant discourse entity under discussion in both the current and previous sentences. C_B^n is the backward-looking center of the n-th sentence of a given discourse.

The forward-looking center list, notated C_F^n for the n-th sentence, is a list of all the discourse entities in a sentence. This list is ordered according to argument role, using the standard hierarchy, sometimes referred to as *grammatical obliqueness*. The *subject* is the least oblique argument, and becomes the first element of the forward-looking center list. By virtue of this privileged position, it is also termed the *preferred center*, C_P^n. The remainder of the forward-looking center list consists of the *direct object*, then *indirect objects*, and then *adjuncts*.

In standard centering there are three transition types, *continue*, *retain* and *shift*, and in BFP shifts are broken down into two subtypes.

Continuing is when the backward-looking center is unchanged ($C_B^{n-1} = C_B^n$), and is also the preferred center of the new sentence ($C_B^n = C_P^n$).

Retaining means the backward-looking center is unchanged ($C_B^{n-1} = C_B^n$), but is no longer in preferred position ($C_B^n \neq C_P^n$), signaling that a shift is likely to occur in the following sentence.

Shifting is what happens when the new backward-looking center is different from the old ($C_B^{n-1} \neq C_B^n$). If the backward-looking center is the same as the preferred center ($C_B^n = C_P^n$), the transition is known as a **smooth** shift. If the backward-looking center is different from the preferred center ($C_B^n \neq C_P^n$) , what results is a **rough** shift.[3]

When resolving anaphora, different analyses may correspond to different transition types. A ranking is given over the different transitions: analysis involving least change (and the least hint of coming change) is preferred. Thus, continuing is favored over retaining, which is preferred over a smooth shift, which in turn is preferred over a rough shift.

The process of anaphora resolution is based on a multi-stage algorithm which involves firstly generating alternative resolutions, then pruning out those resolutions that conflict with certain absolute constraints, and then applying the transition ranking. In more detail, the algorithm runs as follows:

Construct The alternative possibilities for anaphoric resolution are identified. Each possibility maps pronouns in the sentence to dis-

---

[3] BFP use the terminology *shift* for a smooth shift and *shift-1* for a rough shift, but the texture-based terminology is now more common.

course entities in such a way as to respect agreement features. For each possibility, C_F^n consists of all the referents of NPs in the sentence, and C_B^n is chosen from C_F^{n-1}, or is chosen to be NIL. A NIL backward-looking center means that there is no link to a previous sentence, for example in an initial sentence of a discourse.

Filter Possibilities are discarded unless all of the following conditions are met:

1. If there are pronouns in the current sentence, then one of them refers to the backward-looking center of the current sentence;

2. The backward-looking center is mapped onto the entity mentioned in the current sentence which is highest ranked in the previous sentence's forward-looking center list;

3. Syntactic coreference constraints (presumably derived from GB binding theory) are upheld.

The first of these is what is known in the Centering literature, following Grosz et al. (1995), as *Rule 1*.[4]

Classify Classify each possibility as one of the transition types using the criteria above.

Select Choose the best possibility, using the ranking over transition types *continue > retain > smooth shift > rough shift*.

2.3. APPLICATION OF THE BFP MODEL

Consider the treatment of the third sentence in the following example, where the second sentence is assumed to be already interpreted with the resolution indicated using indices:

(1)   a.   Jane$_i$ likes Mary$_j$.

        b.   She$_i$ often brings her$_j$ flowers.

        c.   She chats with the young woman for ages.

For this example the forward-looking center list from the second sentence C_F^2 will be ⟨Jane, Mary, flowers⟩.

---

[4] Rule 2 of Grosz et al. (1995) is a preference relation over sequences of transitions, which is partially captured by the Classify and Select phases of the BFP algorithm. As pointed out by a reviewer of the current paper, the fact that BFP transition preferences are evaluated over pairs of sentences, whereas Rule 2 is supposed to apply over longer sequences may be empirically significant. But it is not clear exactly how to apply Rule 2 over larger sequences: the proposal in 5, below, provides one approach.

Table I. Possible resolutions of example $(1c)^5$

	C_B^{1c}	C_F^{1c}	Filters
1	Jane	$\langle$ <u>Jane</u>, Mary$\rangle$	
2	Jane	$\langle$ <u>Mary</u>, Jane$\rangle$	
3	Mary	$\langle$ <u>Jane</u>, Mary$\rangle$	2
4	Mary	$\langle$ <u>Mary</u>, Jane$\rangle$	2
5	NIL	$\langle$ <u>Jane</u>, Mary$\rangle$	1,2
6	NIL	$\langle$ <u>Mary</u>, Jane$\rangle$	1,2
7	flowers	$\langle$ <u>Jane</u>, Mary$\rangle$	1,2
8	flowers	$\langle$ <u>Mary</u>, Jane$\rangle$	1,2
9	Jane	$\langle$ <u>Jane</u>, Jane$\rangle$	3
10	Jane	$\langle$ <u>Mary</u>, Mary$\rangle$	3
11	Mary	$\langle$ <u>Jane</u>, Jane$\rangle$	2,3
12	Mary	$\langle$ <u>Mary</u>, Mary$\rangle$	2,3
13	NIL	$\langle$ <u>Jane</u>, Jane$\rangle$	1,2,3
14	NIL	$\langle$ <u>Mary</u>, Mary$\rangle$	1,2,3
15	flowers	$\langle$ <u>Jane</u>, Jane$\rangle$	1,2,3
16	flowers	$\langle$ <u>Mary</u>, Mary$\rangle$	1,2,3

Construct Agreement facts prohibit "she" or "the young woman" referring to flowers, so the only possibilities constructed involve each of these expressions referring to Jane or Mary. This results in the 16 possibilities for the pair $\langle C_B^{1c}, C_F^{1c}\rangle$ shown in table (2.3).

Filter Items 5–8 and 13–16 fail the first filter, since the backward-looking center is not pronominalized despite the presence of a pronoun. Items 3–8 and 11–16 fail the second filter, since C_B^n is not the most salient item mentioned, and items 9–16 fail the third, which prohibits argument coreference. This leaves us with items 1 and 2.

Classify Item 1 is classified as *continuing*, since $C_B^{n-1}=C_B^n=C_P^n$. Item 2 is classified as *retaining* since $C_B^{n-1}=C_B^n\neq C_P^n$.

Select Candidate 1 wins out, since continuations are ranked higher than retentions. So it is predicted that "she" refers to Jane, and "the young woman" to Mary.

2.4. OBSERVATIONS

Before moving on to consider the COT reformulation of centering, I would like to point out some peculiarities and shortcomings of BFP.

First, the algorithm takes a parsing/interpretation oriented perspective. Although application of BFP to generation is discussed briefly in

the original paper, the algorithm as described above is not reversible in any obvious way.[6]

Second, the only anaphors dealt with in the published algorithm are pronouns. This, in turn makes the status of the Rule 1 filter peculiar. Rule 1 is normally taken to mitigate against using definite descriptions for C_B, and to prevent interpretation of definite descriptions as coreferential with C_B when a pronoun is also present. However, the lack of definite descriptions in BFP means that such situations do not even arise within the theory's application domain.

What effects, then, does Rule 1 have in BFP? One effect is to filter out pathological possibilities where C_B^n is not even mentioned in the current sentence despite the presence of anaphoric links. In example (1), above, interpretations involving "NIL" and "flowers" failed both the first filter (Rule 1) and the second filter. Such examples of filtering do not seem to correlate with anything that the original architects of Centering might have had in mind as a function for Rule 1, or anything motivated by any explicit empirical study. Besides this, it is notable that *all* the readings filtered by Rule 1 in (1) would also be filtered by the second filter, whereas the reverse is not true. If this type of filtering were the only motivation, Rule 1 would be completely superfluous.

The second effect of Rule 1 relates to its originally intended function, and can be seen in constructed examples where a proper name happens to corefer with the previous preferred center, and agreement prevents the only pronoun in the sentence from referring to the previous preferred center. Unfortunately, such examples do not necessarily support the BFP model. Consider example (2):

(2)   a.   Mary likes tennis.
                $C_B^1 = NIL$, $C_F^1 = \langle Mary,\ tennis \rangle$, *rough shift.*

        b.   She plays Jim quite often.
                $C_B^2 = Mary$, $C_F^2 = \langle Mary,\ Jim \rangle$, *smooth shift.*

        c.   He used to play doubles with Mary.
                $C_B^3 = Jim$, $C_F^3 = \langle Jim,\ another\ Mary \rangle$, *smooth shift.*

The final sentence of (2) is difficult to process. It is comprehensible if "Mary" is de-stressed and nuclear sentence accent is placed on "doubles", although there remains a feeling of incompleteness, perhaps as if the speaker was going on to say something further about *Jim*. On this reading, the uses of "Mary" are coreferential.

According to BFP, the reading of the third sentence where "He" refers to Jim and "Mary" refers to the same individual named in the

---

[6] C.f. Kibble (1999) for an attempt at basing a generation component on the BFP model.

first sentence (and referred to in the second) is predicted not to be available. It is filtered out because Mary would be the backward-looking center but not pronominalized while there is a pronoun present. The BFP prediction that this reading is ruled out is incorrect, although some prediction of processing difficulty would be appropriate. In fact the authors of BFP mention the possibility of making Rule 1 into a preference rather than a constraint.[7] We return to the treatment of (2) shortly.

The possibility of altering the status of Rule 1 brings me onto my next point: given that linguistic generalizations can be expressed at any of the algorithm's four stages, how are we to judge where a particular generalization belongs?

Part of Rule 1 could easily have been expressed in the construction stage. The algorithm could have been altered in such that the only interpretation candidates considered map C_B^n onto some element occurring in both C_F^n and C_F^{n-1}, and only onto NIL if there was no such element. Similarly, other filtering constraints could have been expressed as construction rules, and *vice versa*. Why should a syntactic agreement test be built into the construction phase, but a syntactic co-occurrence test be built into the filtering stage?[8]

On the other hand, the suggestion that Rule 1 be made a default essentially amounts to moving it into the Classification and Selection phase. To make Rule 1 a default in a way consistent with the general framework, it would seem that we would have to double the number of transition types. Each of the current transitions would bifurcate into one version in which Rule 1 was followed, and one in which it was not. Having thus defined eight transition types, a linear ordering would then be defined over them. There are, in principle, $8! = 40320$ such orderings.[9]

---

[7] Whether the model of Grosz et al. (1995) predicts complete absence of the reading of (2c) in which both "Jim" and "Mary" are anaphoric is unclear: at the very least their discussion should lead us to expect that such a reading would create processing difficulties for the hearer. I am grateful to a referee for pointing out that predictions of Brennan et al. (1987) may differ from those of Grosz et al. (1995) in this respect.

[8] The reader is referred to Gordon and Hendrick (1997) for extensive psycholinguistic studies on the interplay of syntactic constraints with other aspects of Centering theory.

[9] In BFP, transition classification is based on two binary constraints, $C_B^n = C_B^{n-1}$, and $C_B^n = C_P^n$. If assume that Rule 1 is more important than these two classification constraints, and that the requirement that $C_B^n = C_B^{n-1}$ continues to take precedence over the requirement that $C_B^n = C_P^n$, we would be left with just one ordering over classifications that incorporated Rule 1. Mere inclusion of Rule 1 into the transition classification schema is thus not particularly difficult, although it would lead to an ordering over eight different transition types. My point is that the use of transi-

More generally, if we have k independent binary classification constraints, we have $2^k!$ possible orderings over transition types. Adding further defeasible classification constraints to BFP produces huge numbers of possible orderings, and there is little reason to believe that this space of orderings will be useful to the linguist or the computational linguist. Rather than discussing methodological and implementational issues which this raises, I now move on to an alternative analysis in which orderings over constraints are defined directly, instead of being defined indirectly via orderings over transition types.

3. Reformulating Centering

In this section, I will introduce relevant aspects of Optimality Theory, present a version of Centering in terms of a simple and motivated set of ranked constraints, show how the model can be applied to examples, and finally discuss the formal relationship between the new model and the BFP Centering account.

3.1. BASICS OF OPTIMALITY THEORY

In OT models, there are standardly two levels of representation, an input and an output. In OT syntax is a representation of LF or argument structure, and the output is a representation of surface structure — see e.g. Grimshaw (1997). Relative to a given fixed input, the constraints are used to find the optimal output.

In the COT model, the two levels of representation are again, roughly, form and meaning. The first is a (partially) syntactically analyzed sentence, and the second is a mapping from referring NPs in the sentence to their referents. From a parsing/interpretation perspective, we take the form as input, and calculate the optimal output. In section 5 it is shown how the system can be used 'in reverse' (i.e., in the direction more normal from the perspective of work in OT syntax) to help select alternative forms on the basis of a fixed meaning.

Given some input to an OT model, the constraints provide a way to select an optimal interpretation from a set of candidates. The set of candidates is assumed to be in principle unrestricted. For instance, the output candidate set in OT syntax could be all the syntactic trees defined over some set of rewrite rules, or the set of all strings over some atomic language. In practice, a given OT paper will generally only consider a set of constraints pertinent to a small group of phenomena,

---

tion classification schemes, like the intuitions behind orderings over them, tends to become rapidly less transparent as the number of classification constraints rises.

and the constraints required to determine other aspects of the input-output mapping are not explicit. It is thus standard to restrict the candidate set to relevant alternatives, those assumed not to be ruled out by constraints that are unrelated to the phenomena at hand. So it will be with COT: the candidate set only partially specifies the meaning of a sentence, and the only candidates that will be considered are those that seem of interest for a theory of anaphora resolution.[10]

Now we come to the question of how the optimal candidate is chosen. Firstly, it should be realized that while some constraints are boolean with respect to candidates, some are not. It is possible that while two candidates both violate a constraint, one candidate involves more violations. Candidate A is superior to candidate B if and only if there is some constraint χ such that (i) A has no more violations than B of each constraint higher ranked than χ, and (ii) A has strictly fewer violations of χ than B.

3.2. CENTERING IN OT: THE BASIC MODEL

Having outlined the basic principles of OT, I now move to the reformulation of Centering. I find it convenient to make a terminological change: *topic* instead of *backward-looking center*. The identification of these two notions is argued for by Ward (1988) and suggests links between Centering and a wide literature based in quite different empirical domains within linguistics.[11] In this section and the next, the term *topic* is

---

[10] One issue which is not dealt with in this paper is the nature of what in OT is called GEN, the function/algorithm that creates the candidate set. I assume that GEN creates pairs of all possible forms and meanings with no further restriction. The limited sets of constraints that are considered mean that COT is only sensitive to a few select features of the forms and meanings, such as the obliqueness of arguments and identity of referents. Some forms or meanings generated may be so unlike what we expect of forms or meanings that features like obliqueness and referent identity are undefined for them. In this case, these forms/meanings are assumed to violate all relevant constraints, thus rendering them non-optimal, and irrelevant to our considerations. For example, GEN might produce a pair consisting of a certain sentence and a peanut: the peanut will be a candidate meaning, but certainly non-optimal. I will omit peanuts and other oddities from the tableaux in this paper.

[11] I first became aware of theorists other than myself having made the terminological shift to using *topic* for C_B in a talk Ellen Prince (relevantly, Ward's thesis advisor) gave at the January 2000 LSA meeting in Chicago. Historically, the decision of the earliest Centering theorists to adopt a novel terminology rather than using existing terms such as *topic* presumably stems from two motivations. First, Centering theory demands several types of *center* operating in concert, whereas *topic* is typically understood to pick out a single linguistic entity. Second, it is by no means clear that linguists use such terms consistently, and there may have been a perceived danger of the new theory being polluted by association with various existing notions in the literature. Note here that linguists often use the term *topic*

merely an abbreviation defined as in (3). The reader could substitute the definiendum for all occurrences of the definiens with no loss of predictive power, but some loss of intuition, supporting motivation and conciseness.

(3)   The *topic* of a sentence is the entity referred to in both the current and the previous sentence, such that the relevant referring expression in the previous sentence was minimally oblique. If there is no such entity, the topic is undefined.[12]

The various generalizations on which the BFP resolution algorithm is based will now be expressed using six linearly ranked constraints. Additionally, I will require that a list is maintained of which entities were referred to in the previous sentences, what the grammatical obliqueness of each referring expression was, and which was topic. No further apparatus specific to Centering is required. Here are the constraints, in rank order, with the top constraint being the highest ranked:

(4)   **AGREE** Anaphoric expressions agree with their antecedents in terms of number and gender.

        DISJOINT Co-arguments of a predicate are disjoint.

        PRO-TOP The topic is pronominalized.

        FAM-DEF Each definite NP is familiar. This means both that the referent is familiar, and that no new information about the referent is provided by the definite.

        COHERE The topic of the current sentence is the topic of the previous one.

        ALIGN The topic is in subject position.

I will now discuss the ordering and the constraints themselves, clarifying the role of each constraint and the extent to which it is independently motivated in the linguistic literature.

The constraints are generally recognizable from BFP, and the ordering is also related to the BFP algorithm. Suppose that two COT

---

to contrast with the term *focus*, while Sidner (1983) uses the term *focus* to pick out various theoretical constructs close to Centering's *forward-looking center*, constructs quite unlike linguists' *focus*. This must have made the adoption of a new terminology attractive. I am not aware of any argument in the literature against the conflation of *topic* and *backward-looking center*. Further discussion of *topic* is postponed until section 4.

[12] To clarify, for a discourse initial sentence the second clause of the definition of topic is intended to apply. This can be compared with the effect of C_B^1 being set to NIL in BFP.

constraints mirror operations taking place in the BFP algorithm, and that the operation corresponding to the first constraint takes place earlier in the algorithm than the operation corresponding to the second constraint. Then the first constraint is higher ranked in COT than the second constraint. There are two exceptions to this principle. First, FAM-DEF does not correspond directly to a BFP constraint. Second, COHERE and ALIGN both relate to a combination of the Classify and Select stages of the BFP algorithm, and their relative ordering is not determined by temporal precedence in the algorithm. Rather, the relative ranking of these two constraints mirrors the ranking of transition types in BFP. Specifically, the ranking reflects the fact that in Centering transitions involving a constant topic (i.e. C_B) are preferred to those where topic changes, independently of whether the topic is in subject position. Examples below will clarify this.

The top two constraints, AGREE and DISJOINT reflect ideas that are familiar from the syntactic literature, the second mirroring the effect of *Principle B* from binding theory.[13] They are found in the construct and filter stages of BFP, respectively. Their relative ordering is arbitrary in the current work.

PRO-TOP has essentially the effect of Centering's Rule 1. However, the original Rule 1 includes an if-clause; "if there are pronouns in the sentence then. . . ." Given that the original rule was an absolute constraint, it was essential that the if-clause restricted the rule's application. However, in COT constraints are defeasible. If there are pronouns, then PRO-TOP will function comparably to Rule 1, providing a preference for interpretations that make the topic (i.e. C_B) into a pronoun. But if there are no pronouns, then all candidate interpretations will be equally bad as far as PRO-TOP is concerned, which means that PRO-TOP will not have any effect in the final preference over candidate interpretations. Similarly, undefinedness of the topic will result in a violation, but in some cases, such as the first sentence of a discourse, that will have no net effect on interpretation. Examples in the following section should clarify.

Motivation for PRO-TOP derives from the cross-linguistically recognized principle that topics are reduced. For example, Bresnan (1999) suggests a similar constraint to PRO-TOP: "Reduced ⇔ TOP". More generally, a rule governing the form of the topic is just a special case of the rules relating the form of NPs to their incoming and outgoing salience. Such rules, at least in as far as they relate form to incom-

---

[13] Statements of Principle B typically differ from DISJOINT, in that there is an explicit caveat clause allowing that arguments may corefer if they are marked as identical by the use of a reflexive. Such a caveat could be replaced in OT by the addition of a highly ranked constraint stating that reflexives corefer with coarguments.

ing salience, have been developed in work on the Givenness Hierarchy (Gundel et al., 1983).

The observation that definites should be familiar, enshrined in FAM-DEF, has a long history in the linguistic literature: Abbott (2001) provides a good summary discussion, and dates the introduction of the idea back at least to Christopherson (1939). Heim (1982) provides a thorough overview and discussion of the familiarity condition, as well as giving her classic formalization of the notion of familiarity using *file-change semantics.*

It should be noted that the class of definites governed by FAM-DEF is taken to include pronouns, definite descriptions and proper names. While it is common to group these (and other expression types) together as definites (c.f. Abbott, 2001), it is not clear that they all share the same familiarity conditions. For current purposes I will make the simplifying assumption that familiarity is a matter of prior mention in the discourse of all the information required by the definite. A more sophisticated account of familiarity might be built, for example, on the work of Prince (1981), who provides a well-known taxonomy in which distinctions are made between speaker new and hearer-new information.[14]

One subtle aspect of OT is that combinations of default rules can have the effect of producing absolute constraints. In this particular case, the interaction of PRO-TOP and the lower ranked FAM-DEF produces the effect of Rule 1's indefeasibility. To show this, it is important first to clarify the interpretation of FAM-DEF.

Suppose that there is a possible interpretation of a sentence where some proper name or definite description refers to the topic. Suppose further that there are pronouns in the current sentence which are anaphoric but refer to discourse entities other than the topic. Then this interpretation cannot be optimal. Why? Because this reading breaks

---

[14] Note that although proper names are preferentially familiar, this will not mean that a use of "Jane" is taken to be anaphoric upon a previous use of "Mary". The notion of familiarity requires that "no new information about the referent is provided by the definite." Thus a use of "Jane" coreferential with a previous use of "Mary" would constitute a violation of FAM-DEF, just as a non-anaphoric use of "Jane" would. In effect, the constraint will cause multiple uses of "Jane" to preferentially refer to the same individual, except where higher ranked constraints say otherwise.

It may clarify to consider a discourse suggested by one of the anonymous reviewers: "Jane_i saw Mary_j. Jane_k was crying." Here, whether "Jane_k is interpreted anaphorically or not, the second sentence will involve a violation of PRO-TOP, either because there is no topic defined or because the topic is not pronominalized. The only constraint that can differentiate between anaphoric and non-anaphoric interpretations is then FAM-DEF, which will lead to a preference for the intuitively correct resolution, $k = i$.

PRO-TOP and not the lower ranked FAM-DEF. But there must be alternative interpretations which break FAM-DEF by allowing the proper name or definite to refer to a novel entity. All these alternatives are such that the topic can be identified with the referent of some pronoun, so they do not conflict with PRO-TOP. Thus they are preferred to the original interpretation which did conflict with PRO-TOP. We will see an example of this reasoning shortly.

The last two constraints, COHERE and ALIGN are just the conditions used in BFP to specify transition types. COHERE, which is satisfied only if there the topic is defined and unchanged, is motivated for example in the work of Givón (1983) and Thompson (1987).[15]

I have been asked several times why, given COHERE, the topic ever changes. From the point of view of interpretation, the answer is simple: the topic changes because COHERE is ranked lower than other COT constraints. From the point of view of production, the answer is more subtle. A speaker can change topic because COHERE is ranked lower than other constraints that here remain unstated, constraints which demand that a discourse compactly conveys information that is useful and relevant to wider communicative goals. When production is considered in section 5, I will simply assume pre-given meanings that must be expressed, even at the expense of violating COHERE, and will limit the set of candidates to those which faithfully express these meanings. This paper is concerned with the relationship between form and meaning, not with the question of what should be expressed in the first place.

ALIGN literally requires the topic to be subject, but for canonical English sentences this is equivalent to saying that the topic is the preferred center of the current sentence. Cross-linguistic motivation for ALIGN can be found in the work of Keenan (1976) and in recent literature on OT-syntax, for instance the treatment of Swedish by Sells (2000).[16]

---

[15] The presence of this constraint requires that an interlocutor's *information state* determines what the topic of the previous sentence was, but I will not explicitly define notions of information state or *update* in this paper. The dynamic model of anaphora resolution of Beaver (1999a) does make explicit a notion of information state — the states used there would have to be augmented with a *topic register* if an interface between COT and Dynamic Semantics were to be developed.

[16] Note that Sells uses a combination of two constraints, one to say that the topic is left aligned in the clause, and another to say that the subject is left aligned. It is only in *canonical* sentences that these produce the same effect as the single ALIGN constraint used here. My ALIGN is sufficient for demonstrating the COT framework, but further work ought to explore the use of constraint combinations to model effects of non-canonical word order on coherence and anaphora resolution. Also note that in other work (Sells, 1999), Sells uses *prominence* relationships which mirror the use

As a further indication of the place of ALIGN and COHERE within linguistic theory, consider the following quote from Katz (1980) (partially cited also by Brown and Yule, 1983): "The surface-subject position imposes the rhetorical or stylistic role of DISCOURSE TOPIC on an NP occupying it, especially one that has been moved into that position.... The notion of a discourse topic is that of the common theme of the previous sentences in the discourse, the topic carried from sentence to sentence as the subject of their predications." Provided we allow that Katz's notion of *imposes* is the defeasible preference found in Centering theory, then it is clear that the backward-looking center is very like what Katz referred to as the discourse topic. Note that my use of *topic* as opposed to *discourse topic* is consistent with much contemporary use of the term in syntactic theory — see e.g. the uses of the term by Aissen (1992), or the even more recent discussion of Chinese of Shi (2000).

The authors of BFP, although they do not share my terminology, make it clear that COHERE is more important than ALIGN. Where we differ is that in BFP the relative ranking of these constraints is stated indirectly, as an ordering over transitions, whereas in COT the ranking is stated directly. More generally, in COT all constraints are ranked directly.

If we wanted to expand BFP to include k defeasible constraints we would have to decide between $2^k!$ transition rankings. In COT, the number of rankings for a given k is $k!$, and increases much more slowly than in BFP. This, of course, is no argument for COT being *a priori* a better model than BFP, or *vice versa*. But it may be suggestive of why, from personal experience, I find direct rankings over constraints easier to work with than rankings over transitions.

3.3. APPLICATION OF COT

In this section, COT is applied to a range of examples. Interpretations of sentences are compared using a tableau method which is standard in OT. To begin, consider the following discourse:

(5)   a.   Jane$_i$ likes Mary$_j$.
        b.   She$_k$ often visits her$_l$ for tea$_m$.
        c.   The woman$_n$ is a compulsive tea drinker.

Since there is no previous sentence, by definition the topic of (5a) is undefined. COT plays no role in the interpretation of this sentence,

---

of the forward-looking center list in centering. In the basic version of COT, these prominence relationships are built into the definition of topic, but I take this to be provisional. See also Aissen's use of scales, e.g. Aissen (1999).

but for the record we can note that there are violations of PRO-TOP, FAM-DEF, COHERE and ALIGN.

For the second sentence, (5b), we proceed by constructing a tableau. In the tableau, (6), the top row lists the input, which here is identified by the relevant example number, and then the constraints in rank order, with the strongest constraint on the left. Each of the following rows details the behavior of one candidate interpretation with respect to the constraints, each star marking a constraint violation. The best of the candidates under consideration is found by looking at each constraint column from the left until finding a column in which one candidate has fewer violations than any other. This is then the optimal candidate, and is dignified with a "☞".

(6)

Example (5b)		AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞	$k = i, l = j$					*	
	$k = l = i$		*			*	
	$k = j, l = i$					*	*
	$k = l = j$		*			*	
	$k = i, l \notin \{i,j\}$				*	*	
	$k = j, l \notin \{i,j\}$				*	*	
	$k \notin \{i,j\}, l = i$				*	*	*
	$k \notin \{i,j\}, l = j$				*	*	*
	$k, l \notin \{i,j\}, k \neq l$			*	**	*	*
	$k = l \notin \{i,j\}$		*	*	**	*	*

I will now consider the role played by each constraint in (6) individually.

1. AGREE: no candidates produce agreement violations.

2. As regards DISJOINT, there are violations whenever the two pronouns, both arguments of the predicate "visits", are resolved to the same entity.

3. The constraint PRO-TOP is violated in two cases, the lowest two, in both of which the topic is undefined because there is no anaphoric reference at all. In these two cases all constraints that make reference to the topic are violated simply because there is no topic.

4. FAM-DEF is violated by the lower six candidates, in all of which at least one of the two pronouns is not interpreted as anaphoric.

5. COHERE is violated in all cases simply because the topic of the previous sentence was not defined.

6. We now come to the issue of how to decide between the only two 'sensible' candidate interpretations, the first and third, in each of which one pronoun picks up the previous subject, and one the object. In either case, the topic of (5b) will be Jane, since Jane is referred to in the least oblique position in (5a). ALIGN mitigates in favor of the parallel reading, producing a violation in case the topic is not in subject position. This gives us the final result: (5b) is predicted to mean that Jane often visits Mary for tea.

The third sentence of (5) is one involving a definite description and no pronoun. I have not imposed any requirement that the definite description actually does describe what it refers to: presumably in a more complete model this would be a high ranking constraint. But the matter is more complex. A full treatment of definites would involve consideration of the extent to which the evolving common ground establishes what a description refers to, and consideration of the ease with which information not yet established can be *accommodated* (in the sense of Lewis, 1979). These issues, although crucial, go beyond what is standardly discussed under the rubric of Centering, and beyond what I aim to achieve in this paper. See Blutner (2000) and Zeevat (1999) for discussion of accommodation in an OT framework.

I will consider four resolution possibilities for the definite NP "The woman$_n$" in (5c), $n = k$, $n = l$, $n = m$, and $n \notin \{l, m, n\}$. PRO-TOP fails for all candidate interpretations, so this constraint does not end up affecting resolution. The $n \notin \{k, l, m\}$ candidate where the definite is not anaphoric fails on three additional counts, since its lack of anaphoricity means that the topic is undefined. The candidate mapping "The woman$_n$" onto tea violates PRO-TOP, FAM-DEF and COHERE: FAM-DEF fails not because the referent is new, but, somewhat bizarrely, because "The woman$_n$" is not already established to be tea.[17] The $n = l$ candidate, where the definite refers to Mary, involves a familiar reference for the definite but not continuity of topic. Thus this candidate loses out to the $n = k$ candidate which maps the definite onto Jane, the topic of the previous sentence. The tableau is in (7).

---

[17] Arguably "The woman$_n$" also differs from "tea$_l$" in grammatical gender, so that the second candidate in (7) also violates AGREE. The question of whether non-pronominal anaphors in English have grammatical or semantic gender is not tackled in the current paper.

(7)

Example (5c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞ $n = k$			*			
$n = l$			*		*	
$n = m$			*	*	*	
$n \notin \{k, l, m\}$			*	*	*	*

Example (5c) illustrates the fact that PRO-TOP, although it is closely related to Rule 1, does not require the extra if-clause "if there are pronouns in the sentence then. . . ." In case there are no pronouns, PRO-TOP simply becomes irrelevant to the choice of candidate.[18]

In the remaining examples in this section, the first and second sentences are assumed to have been processed, resulting in the anaphoric relationships indicated by co-indexation. Examples (8c) and (9c) both involve a pronoun in subject position that can agree with the previous subject. A theory which required that parallelism be maximized would presumably resolve the subject pronoun to the referent of the previous subject in both cases. However, parallelism is not a deciding factor *per se* in COT or BFP, and neither is subjecthood of the antecedent. It happens that according to both models, in (8c) the antecedent is the previous subject. But both models agree that this is not the case for (9c).

(8)   a. Jane_i likes Mary_j.

   b. She_i often goes around for tea with her_j.

   c. She_k chats with the young woman_l for ages.

(9)   a. Jane_i is happy.

   b. Mary_j gave her_i a present_k.

   c. She_l smiled.

The tableau for (8c) is shown in (10). Note that all candidates are here analyzed as violating FAM-DEF since the antecedent of "the young woman_l" is not established to be a young woman in the discourse,

---

[18] Example (5c) is what Centering would classify as a *continuation*, although, from a formal point of view, the BFP algorithm does not cover this particular case: as indicated BFP does not include any explicit treatment of definite descriptions. I have chosen to include some treatment of definites in part because it allows me to provide examples that illustrate the effects of Rule 1/PRO-TOP more transparently than do examples not involving definite descriptions.

although this plays no role in determining the winning candidate. The first two candidates are the obvious two alternative resolutions, and an extra possibility has been included simply to further illustrate the effect of DISJOINT. As can be seen, the high ranking of this constraint means that the third candidate in the table, a reading in which co-arguments corefer, is far from optimal. The second candidate, in which a definite description is resolved to the previous sentence's subject, produces multiple violations: the topic is not pronominalized, and is not aligned with the subject. The parallel-subject reading does not violate any constraints apart from FAM-DEF, and is selected.

(10)

Example (8c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞ k=i, l=j				*		
k=j, l=i			*	*		*
k=i, l=i		*		*		

In (11) a tableau for (9c) is presented. Here the two candidates included are one in which the single anaphor corefers with the previous subject, which was not topic, and one in which the anaphor corefers with the previous direct object, which was topic. The first of these produces a conflict with COHERE, and is ruled out.

(11)

Example (9c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
l=j					*	
☞ l=i						

Modulo the treatment of definite descriptions, it is clear that both (8c) and (9c) would be classified as continuations. The next example, (12c) is one that BFP would classify as *retaining*. In other words, Jane is maintained as the topic, although no longer realized in subject position. As shown in the immediately following tableau, COT also predicts resolution of the pronoun to Jane.

(12)   a. $Jane_i$ is happy.

        b. She_i was congratulated by $Freda_j$,

        c.   and $Mary_k$ gave her_l a $present_m$.

(13)

Example (12c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞ l=i						*
l=j					*	*

The next two cases illustrate BFP shifts. The first, (14) involves a smooth shift in its third sentence, and the second, (16), involves a rough shift in its third sentence. In both cases COT predicts the same shift of topic as BFP. In the first case, the topic shifts because the only interpretations in which both pronouns gain anaphoric readings involve a reference to an entity that was in subject position in the previous sentence but non-topical there. In this case, shown in the tableau in (15), ALIGN comes into play to determine the optimal choice. In the second case, (16c), the only candidates which satisfy COHERE and ALIGN would violate other higher ranked constraints. In the tableau, (17), only the correct reading and another candidate violating AGREE are shown.[19]

(14) a. Jane$_i$ is happy.

b. Mary$_j$ gave her$_i$ a present$_k$.

c. She$_l$ smiled at her$_m$.

(15)

Example (14c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞ l=i, m=j					*	
l=j, m=i					*	*

(16) a. Jane$_i$ is happy.

b. Mary$_j$ gave her$_i$ a present$_k$.

c. Somebody unwrapped it$_l$.

---

[19] I have implicitly restricted the candidate set to rule out interpretations where indefinite NPs are anaphoric. This could, of course, have been stated as a further ranked constraint.

(17)

Example (16c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
l=i	*					*
☞ l=k					*	*

The version of COT under discussion in this section inherits all the empirical virtues and vices of BFP. In particular, the high ranking of PRO-TOP mirrors the centering stipulation that the backward-looking center be pronominalized if there are any pronouns at all. Let us return to (2), repeated below, which was used to illustrate the fact that Rule 1 blocks interpretations which are in fact possible.

(2)   a. $Mary_i$ likes tennis.

        b. She_i plays Jim_j quite often.

        c. He_k used to play doubles with $Mary_l$.

We obtain the following tableau in (18), which shows that the reading where "He_k" is Jim and "$Mary_l$" refers to the same "Mary" as was mentioned earlier is not available. Rather, this version of COT predicts that the preferred reading is one in which we are talking about one Jim, and two different people called "Mary".[20] This situation will be corrected in sections 4.1 and 5.1.

(18)

Example (2c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
k=j, l=i			*			*
☞ k=j, l∉ {i,j}				*	*	
k∉ {i,j}, l=i			*	*		*
k,l∉ {i,j}			*	**	*	*

---

[20] In (18), there are ALIGN and PRO-TOP violations when the topic is not Jim, COHERE violations when the topic is Jim, FAM-DEF violations whenever "He_k"or "$Mary_l$" is used non-anaphorically, and violations of all four when the topic is undefined.

3.4. The Relationship between cot and bfp

The bfp Centering model and the version of cot I have presented up to now are predictively equivalent in a strong, formal sense which I will make clear. To understand this equivalence, it is best to start with a piece of the puzzle, namely how the new proposal can manage without Centering's transitions and yet still produce the same predictions about anaphora resolution.

Having filtered out resolutions that for some reason are considered improper, the final resolution preference in Centering comes down to a competition between alternative transitions. Given that there are four transition types, a total of six pairwise competitions between distinct transition types are possible. These six competitions are shown in the table in (19), each competition separated from the next by a double-line, and the winner of each competition being the topmost of the two. Corresponding to each transition type is a pattern of violations of COHERE and ALIGN, and this pattern is shown on the right-hand side of the table.[21]

(19)

Transition		COHERE	ALIGN
continue	☞		
retain			∗
continue	☞		
smooth shift		∗	
continue	☞		
rough shift		∗	∗
retain	☞		∗
smooth shift		∗	
retain	☞		∗
rough shift		∗	∗
smooth shift	☞	∗	
rough shift		∗	∗

In fact, given that these candidates made it to the transition ranking stage of the bfp algorithm, we know that they satisfy agreement and argument disjointness criteria and have a pronominalized topic. In terms of cot this means that AGREE, DISJOINT and PRO-TOP are satisfied. Restricting consideration to cases in which the only definites are pronouns, and considering only anaphoric interpretations,

---

[21] The various competitions between candidates cited in table (19) are illustrated by (10) and (11) (corresponding to bfp continuing transitions), (13) (retaining), (15) (smooth shift) and (17) (rough shift).

FAM-DEF is also satisfied by both candidates. Thus we know that the preferred candidate in COT will be determined solely by COHERE and ALIGN. As the above table shows, the COT winner in these pairwise competitions is always the same candidate as has a preferred transition in BFP.

Building on considerations like these, a proof of the following proposition is in Appendix A:

(20)   *Given a sentence in which the only definite expressions are proper nouns and pronouns, if either* COT *(with the constraints and constraint ranking above) or* BFP *uniquely predicts an interpretation involving fully anaphoric interpretation of all definites, then both do, and in this case they resolve anaphors identically.*

4. Topic and Salience in COT

The model developed up to now is descriptively faithful to a standard variant of Centering theory. In the remainder of the paper a number of variations and extensions will be considered. In this section I first consider the possibility of the topic being non-pronominal in situations that would violate Centering's Rule 1. I then turn to ways that notions of topic and salience can be encoded as defeasible constraints, rather than being introduced as fixed definitions. As I show, this is advantageous both for the analysis of English and because it provides a locus for the cross-linguistic study of anaphora and text coherence.

4.1. BREAKING RULE 1

COT is flexible in a way that BFP is not: the constraints in COT can be reranked. A simple illustration: suppose PRO-TOP was moved to a position lower than FAM-DEF. In that case the system would make distinct predictions from BFP. There would be a preference for generating anaphoric readings even at the expense of producing what in standard Centering would be a Rule 1 violation. An immediate result of this modification is that the final sentence of (2), repeated below, is correctly predicted to have a preferred reading in which only one "Mary" is referred to, and "He" refers to Jim. This is shown in the tableau (21). This tableau shows the same competition as in the earlier (18), but with a modified ranking of constraints.

(2)   a.   Mary$_i$ likes tennis.

b.   She$_i$ plays Jim$_j$ quite often.

c.   He$_k$ used to play doubles with Mary$_l$.

(21)

Example (2c)	AGREE	DISJOINT	FAM-DEF	PRO-TOP	COHERE	ALIGN
☞ k=j, l=i				*		*
k=j, l∉ {i,j}			*		*	
k∉ {i,j}, l=i			*	*		*
k,l∉ {i,j}			**	*	*	*

As was noted in section 2.4, although example (2) is comprehensible on a purely anaphoric reading of the final sentence, it is awkward. An explanation is given in section 5.1.

The relative reranking of FAM-DEF and PRO-TOP proposed seems reasonable on the strength of this artificial example and others, but I make no strong empirical claim. I leave the hard empirical work, be it based on corpora, research with informants or psycholinguistic studies, for the future. What I do want to claim is that whatever the facts of this matter, COT will provide a convenient framework in which to model them.

4.2. A CONSTRAINT-BASED NOTION OF TOPIC

I will now review how COT might be improved with respect to the definition of topic, which itself implicitly makes use of a notion of *salience*. Currently, the topic is defined as the entity referred to in both the current and the previous sentence such that the relevant referring expression in the previous sentence was minimally oblique. Here *obliqueness* is being used as a practical operationalization of *salience*, or what psychologists might term *activation*.[22] Accordingly, I split the discussion below into one subsection concerning the notion of topic, and one concerning alternative models of salience.

Reinhart (1982) argues against defining topics in terms of given material. She argues that topics are primarily what a sentence is *about*, and that givenness is neither a sufficient nor necessary condition for topicality.[23] Even if one fully accepts the points that Reinhart makes,

---

[22] See Arnold (1998) for an extensive discussion of the relation between activation/salience and topic/focus.

[23] Reinhart's notion of topic matches that used in some but not all contemporary syntactic theory. Aissen (1992) does cite the *aboutness* of the topic as a central

that would not necessarily invalidate the use of *topic* in the current paper. The definition of *topic* used above can be stated instead as a high ranking set of OT constraints which relate the topic to what is mentioned and what is salient. As a result of this move, COT is no longer restricted to any one strict definition of *topic*. The constraints defining *topic* can be violated. In particular, they could, in principle, be violated if the alternative was violation of a higher ranking constraint embodying Reinhart's requirement that the topic is what the sentence is about.

Moving to a constraint based notion of topic, while initially remaining faithful to standard Centering theory, can be achieved by removal of the definition of topic used so far, and addition of the following constraints at the top of the COT ranking:

ONE SENTENCE WINDOW Only discourse entities mentioned in the previous sentence are salient.

ARG SALIENCE One discourse entity is more salient than another if the first was referred to in a less oblique argument position than the second in the same sentence.[24]

UNIQUE TOPIC With respect to any sentence, there is exactly one discourse entity which is the topic of that sentence.

SALIENT TOPIC The topic of a sentence is the most salient discourse entity referred to in that sentence, and undefined if no previously salient entities are referred to.[25]

---

feature, whereas, for example, Shi (2000) explicitly sides against any notion of topic based on aboutness. Note also that Reinhart considers three alternative notions of givenness (or *old information*): *predictability*, *saliency*, and *shared knowledge*. None of these are appropriate as characterizations of topic *qua* C_B, but it could be argued that topic is a special case of *saliency*. In particular, the topic as analyzed here is preferentially the most salient object referred to.

[24] We might consider adding to the definition of ARG SALIENCE that the first discourse entity will be more salient than the second if it occurred in a higher clause. Something like this is implicitly assumed in the analyses of (24) and (25), below. Also, to be more precise, the definition should account for cases where some entity is referred to multiple times in the same sentence, in which case on the current definition it might both out-rank and be out-ranked by some other referent.

[25] A potentially fruitful alternative would be replacing SALIENT TOPIC by a constraint that might be called MAX SALIENT TOPIC: the topic of a sentence is maximally salient and undefined if no previously salient entities are referred to. Use of this constraint would favor reference to a salient object above a less salient one. For example, a single anaphor in the second sentence of a discourse segment would preferentially refer to an individual realized as the subject of the first sentence rather than any other entity mentioned in the first sentence. Centering theory, and BFP specifically, predicts no preference.

The above constraints are related to observations made by Grosz et al. (1995).[26] In particular, UNIQUE TOPIC and ONE SENTENCE WINDOW correspond almost exactly to claims made under the titles "A unique C_B" and "Locality of C_B", respectively (p. 210-211). The constraint ARG SALIENCE is a simplification of a claim, titled "Ranking of C_F", that the C_F list (which I take to be no more than a model of saliency) is "ordered according to a number of factors". They later go on to present evidence that "grammatical role is a major determinant of the ranking on the C_F, with SUBJECT > OBJECT(S) > OTHER"(p. 214). The remaining constraint, SALIENT TOPIC corresponds to a definition they give, identical to the BFP definition, that the "most highly ranked element of $C_F(U_n)$ that is realized in U_{n+1} is the $C_B(U_{n+1})$."(p. 209)

We now have a model in which topic is constrained rather than defined. This has two immediate consequences. First, the above set of four constraints could be ranked lower: see section 5 for a discussion of what effects this would have. Second, as already indicated, if there were a generally agreed on definition of aboutness, we could consider adding a constraint ranked higher than those above requiring that the topic is what the sentence is about. In effect this would mean that topics were primarily what a sentence was about, and only secondarily common themes between sentences. Such an analysis would meet Reinhart's stated objections to defining topic as given material, while still preserving the insight that topics generally are just that.[27]

---

[26] I am grateful to one of the referees for clarifying this point.

[27] There are many further questions to be asked about the notion of topic. What if the topic, in the sense I use it, is changing? Are there then two topics? Since my choice of terminology equates topic with C_B, it is clear that there will just be one topic. One case where multiple topics might occur is during switch of topic, in which case there might be both an *old* or *continuing* topic, and a *new*, *switch* or *contrastive* topic. Thus, for example, what is conventionally *wa-* marked in Japanese might then be equated with the *new* topic, not the *old*. Clearly it would be a mistake to equate Japanese *wa-* marked constituents with the C_B. Again, I leave a more detailed examination of this issue, and comparison with Kuno's use of *topic* (Kuno, 1973), for another occasion. Cross-linguistic work dividing topics into separate categories includes Aissen (1992) and Valduví and Vilkuna (1997). Also see the discussion of contrastive topic of Buring (1999), the excellent cross-linguistic discussion of information structure of Lambrecht (1994) or the shorter and more recent overview of McNally (1998). As mentioned, Ward (1988) is a predecessor to this work in that it identifies topic and C_B, and I should also mention that the notion of *link* developed by Vallduví (1990) (see also Vallduvi (1993) was one of the inspirations behind the current proposal.

4.3. ALTERNATIVE MODELS OF SALIENCE

The question of how the topic is defined is distinct from (although closely related to) the question of what the relative salience of different discourse entities is, i.e. how the forward-looking center list is ordered. It is this latter question to which I will now turn.

There have been several suggestions for how the forward-looking center list should be formed that differ from the model in BFP. For example, it has been suggested (Kuno, 1987; Walker et al., 1994) that for Japanese, NPs marked (by the choice of main verb) as *empathetic* are highly salient, and that *wa-* marked NPs are even more salient. This result could be arrived at simply by adding two extra constraints at the very top of the ranking, in the following order:

SALIENT WA If in the previous sentence discourse entity α was realized by a *wa*-marked form, and discourse entity β was also realized in that sentence, then α is more salient than β.

SALIENT EMPATHY If in the previous sentence discourse entity α was marked as empathetic, and discourse entity β was not, then α is more salient than β.[28]

Similarly, it has been suggested initially for German by Strube and Hahn (1999), and then for English by Strube (1998), that NP form in the previous sentence is a better predictor of salience than argument position. Here *form* refers to the question of whether a discourse entity was realized by a null pronoun, by a regular pronominal form, by a short description, and so on. I will return to the issue of NP form in section 6. For the moment, observe that given a suitable notion of *minimal form*, the generalization could be modeled by using the following constraint ranked above ARG SALIENCE:

SALIENT FORM If in the previous sentence discourse entity α was realized by a more minimal form than discourse entity β, then α is more salient than β.[29]

---

[28] For detailed discussion of what it means for an argument to be marked as *empathetic*, readers are referred once more to the works cited above (Kuno, 1987; Walker et al., 1994). For an alternative view on the significance of *wa*-marking, see Portner and Yabushita (1998).

[29] The constraint SALIENT FORM is *not* to be confused with constraints like Bresnan's "Reduced ⇔ TOP", mentioned above. SALIENT FORM concerns the effect of reducing an expression on its future salience, whereas Bresnan's constraint, like PRO-TOP, concerns the interdependence of the form of the expression on its current salience. In terms of standard Centering, if we are considering the form of an NP in sentence n, then SALIENT FORM concerns the relationship of the form of the

This latter line of work includes suggestions for intra-sentential anaphora, which would require removal of the constraint ONE SENTENCE WINDOW. A more general rule than this would be:

LAST S SALIENCE One discourse entity is more salient than another if the first was referred to in the previous sentence and the second was not.

Such a rule, capturing at least partially the idea that salience declines over time, opens up the possibility not only of a treatment of intrasentential anaphora, and perhaps of the relationship between bound and discourse anaphora, but also of longer distance anaphora, such as occurs with the global focusing mechanisms discussed by Grosz and Sidner (1986).

There are many other respects in which the notion of salience could be changed. Paramount is the need to allow entities other than those referred to by a previous NP to be salient, as made clear in recent work of Eckert and Strube (2000). They show that in a sizeable corpus less than half the anaphoric links were to entities explicitly introduced by NPs, and it is clear that models of *bridging*, propositional anaphora, VP-anaphora and temporal anaphora are all dependent on a far better developed notion of salience than that found in standard centering models or COT.

5. Production

Unlike BFP, the model I have proposed is reversible, and can function equally as a model of comprehension or production.[30] In this section I will first show how production is modeled, so that predictions can be made about preferred linguistic realization. I will then show that the constraint system can be applied not only on a sentence by sentence basis, but also to larger text segments. As a result COT makes clear and testable predictions about the relative coherence of different texts,

NP to C_F^n, whereas Bresnan's constraint and PRO-TOP (and Rule 1) concern the relationship of the NP form to C_F^{n-1}. Put simply: SALIENT FORM implies that being pronominalized makes a referent salient in the future, whereas PRO-TOP concerns the fact that referents which are already salient should be pronominalized. The symmetry of these two requirements is suggestive of the possibility that in future work a synthesis can be achieved whereby the two constraints are explained by a common underlying principle.

[30] For quite different views of the relationship between comprehension and production in OTthan are presented here, the reader is referred to Hendriks and de Hoop (2001) and Zeevat (2000).

predictions that match well with some of the data originally used to motivate Centering theory.

5.1. PRONOMINALIZATION

As discussed, (2c) is awkward. Why?

(2) a.   Mary$_i$ likes tennis.

b.   She$_i$ plays Jim$_j$ quite often.

c.   He$_k$ used to play doubles with Mary$_l$.

The answer, I suggest, is that although (2c) has the desired meaning, someone trying to express that meaning would not choose (2c). Rather, COT predicts that the speaker would choose to pronominalize the topic, Mary. This is made explicit in the production tableau in (22). In this tableau, the input is taken to consist of a combination of linguistic context and a meaning. The meaning involves a predicate *utpdw* ("used to play doubles with") applied to the referents j and i. The candidates are alternative linguistic realizations (in which "u.t.p.d.w." abbreviates the string "used to play doubles with") chosen from a very restricted set. Apart from explaining why (2c) is odd, this also serves to demonstrate how COT can be used as a model of production, or at least as a model of those aspects of production which concern anaphora and coherence.

(22)

Context: (2a,b) Meaning: $utpdw(j,i)$	AGREE	DISJOINT	FAM-DEF	PRO-TOP	COHERE	ALIGN
☞ He$_j$ u.t.p.d.w. her$_i$						*
He$_j$ u.t.p.d.w. Mary$_i$				*		*

Clearly in any expansion of this tableau to include further candidate realizations, (2c) will remain dispreferred, and this explains the infelicity of the sentence. However, note that as the theory stands, in such an expanded tableau we may get a different winner, or no clear winner. For example, the instantiation of COT considered so far places no preference for non-topic NPs to be pronominalized, so "He$_j$" could equally be realized as "Jim$_j$", to give "Jim$_k$ used to play doubles with her$_l$". There are various ways that the model could be altered to modify this behavior, for example by adding a low-ranked constraint preferring pronominalization of any NP referring to a salient entity, or by adding a repeated name penalty. For this latter option there is psycholinguistic motivation — see Gordon and Chan (1995).

5.2. A bidirectional perspective on Rule 1 violations

Violations of Rule 1 are known to occur. One reason is that pronominalizing the topic is not always an option. Poesio et al. (2000) report that naturally occurring examples are often cases where the topic is maintained by a bridging description. For example on their classification, Rule 1 is violated by the second sentence of (23), an example which they drew from a corpus of descriptions of museum objects.

(23) a. A great refinement among armorial signets was to reproduce not only the coat-of-arms but the correct tinctures;

  b. they were repeated in colour on the reverse side

  c. and the crystal would then be set in the gold bezel.

Sentence (23b) is considered a violation of Rule 1 because the topic of the first sentence ("armorial signets") is continued using the bridging description "the reverse side", while a less salient entity ("the coat-of-arms") is pronominalized as "they".

Although I have not analyzed plural pronouns or bridging descriptions, this type of example provides support for the analysis I have given. From a production perspective, it is obvious why the NP which Poesio et al. count as C_B of (23b) is not pronominalized. The reference of the bridging description could not be recovered from a pronoun alone, and there is no alternative candidate in which the topic is pronominalized. So the winning candidate is one in which the topic is a full description. The hearer is not only able to derive the correct meaning, but can see no obvious way that the speaker could have expressed it better.

Such examples support the use of PRO-TOP over Rule 1. According to Rule 1, the presence of a non-topical pronoun should produce processing difficulty in (23b), but the sentence is felicitous and easily understood. Rule 1 predicts that the sentence could be improved if "they" were replaced by a description, to give "the coat-of-arms (and the correct tinctures) were repeated in colour on the reverse side." However, this prediction is incorrect. Although the latter version is felicitous, it is not an improvement on the original. PRO-TOP makes the correct prediction here: it penalizes both the original sentence and the new version equally for not having a pronominal topic. Whichever version is produced, it remains the case that there is no way (given other communicative goals which defined the input meaning) for the speaker to avoid violating PRO-TOP. So neither version of the sentence is predicted to be infelicitous.

5.3. The role of synonymy in a bidirectional grammar

More generally, might a speaker choose to produce a sentence that is sub-optimal according to COT, where by *sub-optimal* I mean that there is a better candidate available? Modulo the crucial issue of what processing capacity is available to the speaker, I suggest that the answer is *no*. However, we have to be careful when deciding whether a sentence is sub-optimal. To know that a candidate is sub-optimal from a production perspective is to know that it would be beaten by another candidate. But this depends on knowing what the candidate set is, and the candidate set consists of alternatives that realize the same input meaning, or at least come close enough to realizing that meaning for the speaker's purposes. So to know that one candidate would beat another is to know that the two candidates are synonymous in the given context, or would be viewed as such by the speaker.

The question of whether alternative surface forms have the same meaning is a complex one. Some, most notably Dwight Bolinger, have argued that all distinctions of surface form signal distinct information content.[31] This is a view that is attractive within OT grammar, and perhaps even inevitable. My own view is a slightly weakened variant of Bolinger's: any distinction of form can signal a distinction of information content, but speakers do not always intend to signal such distinctions.

I would like to take a broad view of meaning, encompassing not only literal content, but also implicatures such as those arising from discourse structure. If a hearer can identify what is being signaled by a sentence on a particular occasion of utterance, then I want to define that as a felicitous use of the sentence (which is not to say that it will not feel *marked* or hard to process). But if hearers are unable to identify what is being signaled, they will perceive infelicity. The linguist's starring of a marked sentence would then indicate that there are no contexts in which that sentence could be used to signal something, or at least that the linguist has insufficient imagination to identify an appropriate pair of a context and something being signaled. On this view, the linguist's starring of sentences is a potentially dangerous idealization: it is utterances (or potential utterances) that are (or would be) infelicitous, not sentences.

---

[31] The idea that distinctions in form correspond to distinctions in meaning pervades all of Bolinger's work. This passage from Bolinger (1977) is in many ways typical: "Tell [a man in the street] that if two ways of saying something differ in their words or their arrangement they will also differ in meaning, and he will show as much surprise as if you told him that walking in the rain is conducive to getting wet. Only a scientist can wrap himself up in enough sophistication to keep dry under these circumstances."

For example, it might be argued that COT incorrectly predicts the variant of (2) ending with (2c′) to be infelicitous, because the speaker could have used (2c″) or (2c‴). All three seem truth-conditionally synonymous, but in the latter two cases the topic is realized in subject rather than object position, so avoiding a violation of ALIGN.

(2)   c′.   He used to play doubles with her.

        c″.   She used to play doubles with him.

        c‴. They used to play doubles (with each other).

There is at least one sense in which c′, c″ and c‴ are not synonymous: they give different orderings of the salience hierarchy. One function of a sentence is to prepare the reader for what follows, so a speaker intending to continue (2) by talking about Jim might choose to put the expression referring to Jim in subject position. Thus the sentence "He used to play doubles with her." has functions not achieved by "She used to play doubles with him.", and is correctly predicted to be felicitous.

I will not formalize the argument for why a speaker might choose to violate ALIGN in the case of (2c′). But the issue I will now turn to is closely related: what is at stake is not only how well a sentence fits in its prior context, but also the functional role that a sentence plays in establishing future context.

5.4. TEXT COHERENCE

It is possible to apply COT to compare the felicity of arbitrarily large texts. The only obstacle to doing this is that it is necessary to decide how to count violations of constraints in different sentences. To demonstrate the possibility of optimizing entire texts, I propose that we count violations in a multi-sentence discourse in the most obvious way: we form one tableau using the standard COT constraint ranking, we enter violations of each constraint in the column corresponding to the violated constraint regardless of the sentence in which the violation occurred, and then select the optimal candidate using the standard OT method.[32] On this basis, we can compare, for instance, the two texts of Grosz and Sidner (1998), originally adapted from Grosz et al. (1995), in (24) and (25). The boxes with which I have decorated certain phrases designate which NP would refer to C_B in BFP, or the topic in the basic

---

[32] Grosz et al. (1995) consider incremental online processing of texts, whereas in this section I investigate the possibility of analyzing complete text segments. Note that the issue of whether the theory could be implemented incrementally is distinct from the question of whether the theory makes the right predictions about the coherence of complete text segments. I discuss only the last of these.

COT model from section 3, and the significance of the underlining will
be explained shortly.

(24)   a.   ⬚John⬚ went to his favorite music store to buy a piano.

        b.   ⬚He⬚ had frequented the store for many years.

        c.   ⬚He⬚ was excited to be going to the store to actually buy
              a piano.

        d.   ⬚It⬚ was the biggest music store in the area.

        e.   ⬚It⬚ had just the kind of piano that he wanted.

        f.   ⬚It⬚ was closing just as John arrived.

(25)   a.   ⬚John⬚ went to his favorite music store to buy a piano.

        b.   I̲t̲ was a store ⬚John⬚ had frequented for many years.

        c.   H̲e̲ was excited to be going to ⬚the store⬚ to actually buy
              a piano.

        d.   ⬚It⬚ was the biggest music store in the area.

        e.   He knew that ⬚it⬚ had just the kind of piano that he wanted.

        f.   I̲t̲ was closing just as ⬚John⬚ arrived.

A theory of Centering should account for the fact that (25) is con-
siderably more awkward than (24). However, Grosz and Sidner (1998)
note that the sentence by sentence classifications of transitions in BFP
and indeed the bulk of later Centering literature do not provide any
way to evaluate the coherence of complete texts.

Applying COT to these texts produces the tableau in (26).[33] Rather
than notating violations with the usual "*", I have notated them with
the letter for the line in which the violation occurs, and omitted vi-
olations in the first line, with regard to which the two texts do not
differ.

---

[33] I have taken the two texts to be the only two candidates, and glossed over the
fact that they do not in fact mean the same thing. I assume that all aspects of
meaning where the texts differ are sufficiently insignificant that it is not important
for the text to remain faithful to them. Also note that in (24e) and (25e) pronouns are
used which refer to an individual (John) not mentioned in the previous line, although
this is not reflected in (26). Significantly, this use of a pronoun is felicitous. Perhaps
this indicates that COT should be augmented with a notion of global focus, in the
manner of Grosz and Sidner (1986), or actor focus in the sense of Sidner (1983).

(26)

Common meaning of (24)/(25)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN
☞　　Text (24)					d	
Text (25)			b c f		c f	b c e f

The preference for the first text is correctly predicted, but a few comments on this example of text-optimization are in order.

First, it should be noted that BFP does differentiate between the two texts, albeit crudely. BFP would analyze the second text as involving multiple violations of Rule 1. For instance, in the second line, the use of the full name "John" to pick out the preferred center from the previous line, in combination with the fact that a pronoun is also present in the second line, would be such a violation. Thus the desired interpretation would be filtered out: BFP does not even allow us to interpret (25b) such that "John" refers to the same individual mentioned in the previous line. Similarly, "John" in (25f) could not refer to the same individual as that mentioned in (25e), and allowing BFP to treat definite descriptions in the obvious way, the store mentioned in (25c) could not be the same one mentioned in (25). Whether BFP predicts (25) to be uninterpretable, or whether it predicts that the interpretation involves multiple Johns and multiple stores is not entirely clear, although I have assumed the latter in this paper. Either way, (25) is differentiated from (24), and in a way that suggests explanations for the awkwardness of (25). We may then ask whether BFP provides us with a general way of differentiating good texts from bad. The answer to this question must be negative. Violations of Rule 1 take on an enormous significance in BFP, and can allow texts to be differentiated. But preferences among transitions act at a sentence by sentence level, so that the BFP algorithm does not directly provide a metric for comparing pairs of texts that differ only in terms of which transitions occur. In contrast, COT does provide a simple way of comparing such pairs.

My second comment in regard to the COT analysis of (24) and (25) relates to the fact that the reasons why COT predicts relative infelicity of (25) are arguably different from those cited by Grosz and Sidner (1998). Grosz and Sidner contend that the jerkiness of (25) results from repeated changes in what they term the "center", by which I take it they mean C_B. This seems intuitively reasonable. However, according to the original definition of topic in COT (or of C_B in BFP) the topic changes only twice, and at relatively well spaced intervals, in line c and f. According to COT, the biggest problem with (25) is that it involves

multiple violations of PRO-TOP. In ʙꜰᴘ these would correspond to violations of Rule 1, so that the natural interpretation of the text would not even be available. How can this difference between explanations of infelicity itself be explained?

One possibility is that when Grosz and Sidner were analyzing the two texts, they identified C_B so as to be consistent with Rule 1 wherever possible. Thus, for example, they might have taken C_B of (25b) to be the store, whereas the definitions used in ʙꜰᴘ and ᴄᴏᴛ identify topic as John. This effect could be mirrored in the variant of ᴄᴏᴛ introduced in section 4 by lowering the ranking of the four topic/salience constraints ONE SENTENCE WINDOW, ARG SALIENCE, UNIQUE TOPIC and SALIENT TOPIC. Let us follow the earlier suggestion from section 4 that PRO-TOP is to be below FAM-DEF. I will now consider the analysis of (24) and (25) with the four topic/salience constraints immediately below PRO-TOP.

The new ranking does not significantly affect the analysis of (24), for which, after the first sentence, there were no violations of PRO-TOP, FAM-DEF or any stronger constraint. However, it does affect (25), altering which NPs are taken to be topical. The underlining in (25) marks the principal NP referring to the topic under the new analysis, but the underlining is omitted where the analysis of topic is unchanged. The resulting tableau, from which I have omitted the AGREE and DISJOINT constraints, is presented in (27):

(27)

	Common meaning of (24)/(25)	FAM-DEF	PRO-TOP	O-S-W	ARG-SAL	UNIQUE	SAL-TOP	COHERE	ALIGN
☞	Text (24)			e				d	
	Text (25)			e			b c	b c d	e

For (25), there are no longer any violations of PRO-TOP, but there are two violations of SALIENT-TOPIC, and three of COHERE. On this analysis, in the first four lines, the topic changes from John, to the store, to John, and back to the store, while in none of these cases is the topic shift signaled by violations of ALIGN. This is certainly a sequence of transitions that might justify Grosz and Sidner's informal description. Furthermore, my intuition is that the first four sentences of (25b) are indeed more "jerky" than the last two sentences, lending some further credence to the proposed modification.

The alternative analysis of (25) demonstrates the power and flexibility of the ᴄᴏᴛ framework, and the potential it has for capturing

intuitions about constraints on discourse cleanly. However, it would be foolish to identify the correct ranking of constraints solely on the basis of my intuitions about Grosz and Sidner's intuitions about a single text, and I have provided no evidence at all concerning the ranking of some constraints, such as ONE-SENTENCE-WINDOW. If in addition to reranking all the topic/salience constraints, ONE SENTENCE WINDOW were modified in an appropriate way, then the topic of line (e) might turn out to be John, not the store. Making such a change, and keeping PRO-TOP as the highest ranked constraint governing topic choice, would result in an analysis of the text in which every line had a different topic from the previous one, which is perhaps the analysis that Grosz and Sidner had in mind. It is clear that more refined empirical study is needed.

6.   A Speculative Proposal

In this section I will analyze switch reference of stressed pronouns as a case of *partial blocking* (Kiparsky, 1983; McCawley, 1978), and show how bidirectional optimization in COT might be used to formalize the analysis.

   Above we saw cases where a form was interpretable but the meaning would not be optimal in production. Now we consider the reverse side of this coin, generation of forms which would be incorrectly interpreted. To exemplify this, suppose that in the context of (28a,b), the speaker wished to say that Jim winked, which might potentially be done using any of c, c′ or c″.

(28)   a. Fred_i was eating.

       b. He_i saw Jim_j.

       c. He_k winked.

       c′. Jim_k winked.

       c″. HE_k winked.[34]

---

[34] I use capitalization to indicate that "HE" carries focal stress. In fact, a realization of (28c″), in a context in which winking is not especially salient, would normally include stress on both "HE" and "winked", and optionally each word might be realized as an entire intonational phrase, or an intermediate phrase in the sense developed by Pierrehumbert and used in the ToBI system of intonational transcription — see e.g. Pierrehumbert and Hirschberg (1990). A possible transcription in the ToBI system would involve H*L on the pronouns, and H*L-L% on the verb. However, I skate over details of the intonation for the remainder of this paper.

Let us augment COT with Schwarzschild's constraint AvoidF, "avoid focus" (Schwarzschild, 1999), and rank it below all other constraints. It is easy to see that (28c) will (still) be the preferred realization, as shown in (29).

(29)

Context: (28a,b) Meaning: *winked(j)*	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN	AvoidF
☞ He winked.					*		
HE winked.					*		*
Jim winked.			*		*		

However, (28c), "He winked", is preferentially interpreted as meaning that Fred winked in this context:

(30)

Context: (28a,b) Example: (28c)	AGREE	DISJOINT	PRO-TOP	FAM-DEF	COHERE	ALIGN	AvoidF
☞ k = i							
k = j					*		

So "He winked" is here the optimal form for the meaning *winked(j)*, but *winked(j)* is not the optimal meaning for "He winked".

An ideal speaker should choose a form under the constraint that the optimal interpretation of that form will be the original meaning. This is what Smolensky, in unpublished work (Smolensky, 1998), has termed *recoverability*. Blutner and others (Blutner, 2000; Blutner and Jäger, 1999; Dekker and van Rooy, 2000; Zeevat, 1999) have proposed explaining a wide range of phenomena using a similar approach. These phenomena include *blocking* (Aronoff, 1976): what happens when an apparently sure-fire candidate is beaten in a run-off with a surprise outsider. For example, the sure-fire candidate could be a form produced by regular grammatical processes, and the outsider a conventionalized, perhaps lexicalized, irregular form. Partial blocking occurs when the defeated candidate goes on to enter and win a new contest, in which it would not even taken part had it one its first campaign. Thus a form produced by regular grammatical processes may take on a meaning that would otherwise be secondary or unavailable. The possibility I

will investigate now is that optimality of "he" leads to partial blocking of "HE", such that "HE" takes on a new meaning.

Existing formalizations of bidirectional optimization utilize meta-level mechanisms that sit outside of the standard tableau. However, for current purposes, it is sufficient to illustrate using an unusual constraint within the tableau. This constraint will favor recoverability of meanings, but also the converse, which might be termed *re-generability*: if a form is optimally interpreted as having some meaning, then that meaning should optimally be realized by the original form.

Let us term our new constraint ⋆BLOCK. Given some pool of forms and meanings, a form-meaning pair will violate ⋆BLOCK if it is dominated by (i.e. loses out to) another form-meaning pair in either direction of optimization. We must be wary of the fact that ⋆BLOCK makes reference to what is optimal in the system, which leads to the potential for circularity. To avoid such problems, I define ⋆BLOCK as follows:

⋆**BLOCK** A form-meaning pair may not be dominated by another form-meaning pair in either direction of optimization in the tableau consisting of all constraints except ⋆BLOCK.

It is easiest to see how the analysis works in what I will call a *bidirectional tableau*. A bidirectional tableau allows evaluation of form-meaning pairs, and not just of forms given a meaning, or of meanings relative to a form.[35] Thus we must pick both a candidate set of forms, and a candidate set of meanings. The standard constraints, are then evaluated as normal.

Violations of ⋆BLOCK are easily evaluated. Begin with a bidirectional tableau in which the ⋆BLOCK column is blank, but other constraint violations have been marked as normal. Find any form-meaning pairs in this tableau such that the form is optimal for the meaning but not *vice versa*, or the meaning is optimal for the form but not *vice versa*. Mark these pairs with a star in the ⋆BLOCK column. The optimal pairs are then all those in the resulting tableau for which the form derives the meaning *and* the meaning derives the form.

The following tableau is obtained with ⋆BLOCK ranked below AGREE and DISJOINT, but above everything else:[36]

---

[35] Blutner and others have used variant tableaux to capture non-standard notions of optimality. The ideal bidirectional tableau would be three dimensional, with meanings, forms and constraints each on separate axes, but so far no ideal way of representing this on paper has been found.

[36] I use the victory symbol (☝) to mark that a candidate is both a winning form relative to the meaning, and a winning meaning relative to the form. Note that ☝ is merely an abbreviated way of notating that the candidate is the winner in two standard OT tableaux, one a production tableau and the other a comprehension tableau.

(31)

Context: (28a,b) Meaning	Form	AGREE	DISJOINT	*BLOCK	PRO-TOP	FAM-DEF	COHERE	ALIGN	AvoidF
☞ *winked(f)*	He winked								
	HE winked.			*					*
	Jim winked.				*	*			
winked(j)	He winked			*			*		
☞	HE winked.						*		*
	Jim winked.				*		*		

There is a great deal to be said about this analysis. To start with, "Jim winked" is predicted to be dis-preferred. In this regard, it should be noted that if in some context, like that of a written text, the accent on the pronoun was not readily detectable, then "Jim winked" would indeed become the preferred candidate of those listed. Thus an alternation between "HE" and "Jim" can be predicted. One way of deriving such an alternation formally would be to allow the relative ordering of AvoidF and PRO-TOP to vary. More generally, variations on the available constraints, and variations on the intended meaning, will obviously produce different realizations. For example, a meaning incorporating some implied contrast between Fred and Jim might, with appropriate constraints, be realized using a stressed proper name, as "JIM winked".

I do not want to suggest that the analysis *solves* the problem of focussed (or *strong*) pronouns. I do claim that *BLOCK, and bidirectional OT more generally, provide an interesting perspective on the notion of markedness. The analysis of stressed pronouns is just one example of the more general insight of Blutner and others that bidirectional OT can make predictions which adhere to Horn's *division of pragmatic labour*: "The use of a marked (relatively complex and/or prolix) expression when a corresponding unmarked (simpler, less 'effortful') alternate expression is available tends to be interpreted as conveying a marked message (one which the unmarked alternative would not or could not have conveyed)."(Horn, 1984, p.22). In this case, the marked construction involves special use of accent, and the marked interpretation is one involving a topic shift.[37]

---

[37] The question arises of how the account I have given could be extended to cover other uses of focus. I tentatively suggest that we might draw again on the work of Schwarzschild (1999). He argues that non-focus-marked material is preferentially old, and I propose a more general preference for material to be old (which will compete the with the speakers' need to achieve communicative goals). Suppose

Analyses of stressed pronouns in Centering have previously been given by Kameyama (1999) and Cahn (1995). Cahn's short paper provides some interesting suggestions as to how Centering theory can be combined with the Pierrehumbert and Hirschberg (1990) theory of intonational meaning. The fact that Cahn considers an array of different accent types is a great strength of that paper, and future work on accenting of pronouns should clearly follow that lead — see also the extensive empirical studies of Nakatani (1997).

As was mentioned earlier in this paper, Kameyama's approach to Centering has much in common with COT. And, specifically with regard to accenting of pronouns, her conclusions are also in tune with the proposal I have made. Kameyama suggests a general principle, as follows: "Complementary Preference Hypothesis: A focused pronoun takes the complementary preference of the unstressed counterpart." It is clear that in the case of a single stressed pronoun, Kameyama's principle may fall out from the more general *BLOCK constraint applied here.

One of the most significant restrictions in Centering theory is that it does not provide a sufficiently general account of the form of referring expressions. On the other hand, the so-called Givenness Hierarchy (Gundel et al., 1983) provides a much more general account of the form of referring expressions, but does not attain the degree of precision found in models of Centering such as BFP or COT. The Givenness Hierarchy organizes referring expressions as regards the extent to which they depend on linguistic context for their interpretation, and predicts

---

further that givenness of an expression were defined as an absence of salient alternatives. An absence of focus on an expression would allow that expression to be given its unmarked interpretation, which in turn would mean that a hearer should not bother to calculate what alternative expressions the speaker could have used for that expression. On the other hand, consider an expression with focus. Use of *BLOCK will lead the hearer to determine that the expression has a marked meaning. Exactly what marked meanings are available will depend on the particular expression, but in many cases the least marked alternative will be that the expression is not old. And if it is not old, then there will be salient alternatives. And if there are salient alternatives, then the bidirectional optimization that the hearer performs must take that into account. Thus focus will come to be seen as an invitation to the hearer to consider alternatives, and in considering why the speaker did not use those alternatives, the hearer may draw various inferences. Clearly if I am on the right track then focus triggers a complex (possibly multi-stage) optimization going beyond that described in my earlier analysis of stressed pronouns. I will not attempt to detail this optimization here, but instead give one example: if "I have a DOG" was uttered in certain contexts, the hearer would (i) spot that DOG must have a marked meaning, (ii) conclude that it is not old information, (iii) consider what alternatives the speaker could have used, (iv) recognize that the speaker made a choice not to mention other pets, and then (v) draw the conclusion that the speaker has no other pets.

that speakers always select the most contextually dependent (i.e. given) form that is interpretable by the hearer.

As Zeevat has observed (Zeevat, 1999), the Givenness Hierarchy is naturally formulated in terms of bidirectional optimality. I refrain from a detailed analysis here, but in essence the idea can be communicated in terms of an example.

Suppose that linguistic context makes some discourse entity mildly salient although there are many more salient entities of similar semantic category, and that a speaker must decide between a pronoun, a short definite description and a longer one. In this case, *BLOCK rules out the pronoun because a hearer would not be able to recover the correct meaning. If both the short and long descriptions might potentially lead to recoverable meanings, then the longer can be ruled out by assumption of its inherently greater markedness. Furthermore, if a speaker chooses to use a long description where a shorter one might have done, *BLOCK will ensure that the hearer concludes that a 'special' meaning is intended. This special meaning might, for example, involve the introduction of a new discourse entity (perhaps *accommodation* of a referent in the sense of Lewis (1979)), or it might involve breaking the assumption that the speaker only wished to convey one piece of information, the main predication. Perhaps the speaker wished to also convey certain extra information pertaining to the already mildly salient entity.

The above analysis of accented pronouns is a special case of a wider analysis which remains to be developed in detail, one which might provide hope for a formal combination of Centering Theory and the Givenness Hierarchy. Gundel (1998) provides an excellent discussion of the potential benefits of such an integration.

7. Discussion

7.1. What we have learned about Centering

Until now it was not obvious how the four stage BFP algorithm could naturally be stated declaratively. COT is the first such statement. This declarativity means that COT is equally suited for generation or interpretation. In contrast, the BFP algorithm is suited for interpretation only. It could not be used to generate texts directly: at best it could be used as a filter to determine whether previously generated texts produced the intended interpretation. Generation in COT is a more subtle affair, since COT does not merely filter out texts which lack the desired interpretation. It also filters out texts which capture the the correct interpretation, but capture it sub-optimally.

Another issue which is clarified in COT is the relation between Rule 1 and the two transition classification conditions. In previous work, Rule 1 was seen as qualitatively different from the transition classification tests, despite the fact that no empirical evidence has been cited showing that they are different in kind. In COT Rule 1 is no longer *qualitatively* different from the transition classification tests. All three are stated as defeasible constraints. However, the COT ranking makes Rule 1 *quantitatively* different from the transition classification tests, i.e. stronger. Yet in COT the relative strength of constraints can be altered, and this flexibility is original to the COT framework. It applies not only to the status of Rule 1, but also to other components of the theory, such as the definition of C_B, or topic as I have termed it.

It is not only the internal relationship between the different parts of the BFP model that is clarified in COT, but also their external motivation. As I have made clear, all the constraints used in COThave precedents, generally in research developed independently of Centering theory.

7.2. PROCESSING COSTS

COT provides a perspective on how costs and benefits of various linguistic forms are weighed by the conversational participants. One of the driving forces behind early the Centering proposals of Joshi and associates (Joshi and Kuhn, 1979; Joshi and Weinstein, 1981) was the idea that speakers choose forms which minimize processing costs to hearers.[38] This idea is visible not only in the analysis of accented pronouns proposed above, but also in the analysis of generation and text optimization: COT models the fact that it may be cheaper in the long-run to use an form which is in the short-term relatively expensive. For instance, a speaker may choose a form in which the topic is not in subject position because it will reduce the costs incurred by a *following* sentence in which a topic shift is needed. This idea is explicit in early work on Centering, but submerged in BFP and much following literature: it is formally explicit for perhaps the first time in COT.

---

[38] The clearest presentation of the relationship between costs/benefits and the Blutner-style analysis is found in the work of Dekker and van Rooy (2000). They show that the various non-standard notions of optimality developed by Blutner and Jäger can be viewed in terms of game theory. In this model, relative payoffs of different actions, such as production of a particular linguistic form, are derived from the underlying OT constraint set, and optimality is reduced to special case of strategic equilibrium.

Here I would note that in some recent interpretation directed work on Centering[39], there has been discussion of processing issues. In particular, Kehler (1997) has observed that the speaker's tendency to use computationally cheap shortcuts, like identifying the subject with the most topical discourse entity, is not in fact captured in the BFP model. Strube (1998) has gone further, suggesting replacement of the BFP algorithm with an entirely incremental algorithm which works from left to right through a sentence interpreting each anaphoric expression as the most salient entity discourse entity possible.

One thing Strube's model has in common with COT is that it does away with Centering's transitions. Of course, Strube's model also does away with Centering's predictions, which is one respect in which it differs from COT. None the less, if Strube is right about the efficacy of incremental interpretation, then this still does not show that other facets of Centering should be dispensed with altogether. What it does show is that a theory of discourse, whether it be applied to interpretation or generation, should take account of the processing advantages of incrementally interpretable text. The framework I have described allows processing benefits for the hearer to be reflected in choices made by the speaker. My hope is that this quality will appeal both to those who stick by Centering orthodoxy, and to radicals like Strube.[40]

7.3. DYNAMIC SEMANTICS

In previous work (Beaver, 1999b; Beaver, 1999a; Beaver, 2000) I have developed a framework termed Transition Preference Pragmatics (TPP). This is a proposal for how Dynamic Semantics[41] should interact with pragmatics. One observation motivating TPP is that many proposals

---

[39] A generation perspective was taken in all the early papers on Centering, but the interpretation perspective is dominant in more computational work following BFP.

[40] There is an extensive psychological literature on Centering — see (Hudson-D'Zmura, 1989; Gordon and Chan, 1995; Gordon et al., 1993; Brennan, 1995) or various papers in (Walker et al., 1998). The above discussion of processing factors leads to the question of what the significance of COT is for psychological models. This should be explored in terms of two sub-questions. First, can previous work on preferences for different Centering transitions be reinterpreted in terms of underlying constraints such as I have proposed? Second, can psychologically motivated models of processing cost be used to derive constraints that should be part of a future COT model, or be used to help choose between alternative possible families of constraints? These are not questions which have easy answers.

[41] Here I use Dynamic Semantics in the sense of Groenendijk and Stokhof (1991) and Groenendijk et al. (1995). This is arguably closer to Heim's earlier work (Heim, 1982) than Kamp's (Kamp and Reyle, 1993), although the differences are not necessarily of empirical significance.

in Dynamic Semantics fail to take the process of anaphora resolution sufficiently seriously. What makes this situation acute is that the analysis of anaphora is one of Dynamic Semantics' main applications. The conclusion I argue for is that interpretation of a sentence should not deterministically fix the effect of an information update. Rather, the meaning of a sentence should define a non-deterministic relation between possible incoming linguistic contexts, and possible outgoing linguistic contexts. In TPP, pragmatics provides a preference ordering over alternative incoming-outgoing context pairs. In this way, compositional semantics may underspecify the effect of an anaphor, and the pragmatic component can resolve the underspecification. The model of Beaver (1999a) shows how a simple account of anaphora resolution based on parallelism can be applied in the TPP framework. The current paper develops a richer model of pragmatic interpretation preferences that could, in principle, be interfaced with the TPP semantic component. The result would be a model which incorporated a dynamic notion of meaning, and both semantic and pragmatic constraints on anaphoric linkage. This would produce a system empirically superior either to standard accounts of anaphora in Dynamic Semantics or Centering Theory.[42]

7.4. COT AND OT

Let me now turn to the question of how the account I have presented relates to other work in OT. De Hoop and Hendriks 2001 provide a perspective on the interpretation of quantification and comparatives, and the interaction of these phenomena with (intonational) focus.[43] In the their model, forms are inputs, and meanings are outputs. They go further and suggest that in general OT syntax is speaker based, but OT semantics is hearer based. There is some truth to this: a particular theory is concerned primarily with constraints that refer to only one

---

[42] Roberts (1998) has described a way in which Centering could be integrated with Kamp's DRT. Her goals are closely related to mine. Centering is an intrinsically dynamic theory Yet those with a dynamic bent, who are reading the current paper will be acutely aware that, as noted previously, I say little explicitly about the dynamics of linguistic context, and never specify the details of incoming and outgoing contexts formally. Thus there is much work to be done. A natural way to proceed would be to follow the suggestions of Blutner (2000), who defines preferences over pairs of linguistic forms and output contexts relative to a fixed input context. The proposal in TPP is interestingly related: there preferences are defined over pairs of input-output contexts relative to a fixed linguistic form. This casual comparison suggests that we might eventually consider defining preferences over triples of input contexts *and* linguistic forms *and* output contexts.

[43] The papers (de Hoop, 2000; de Hoop and de Swart, 1998; van der Does and de Hoop, 1998) are on related themes.

grammar component, it is clear that the best demonstrations of that theory will involve taking that component to be the output. However, I would like to suggest that the most important challenge for both OT-syntacticians and OT-semanticists lies in stating the theory that relates these components. A theory that is rich in such relational constraints is not exclusively uni-directional in the way that de Hoop and Hendriks suggest. In COT, all constraints are relational, and the theory can be applied in either direction.[44] OT is suggestive of a perspective on the syntax-semantics-pragmatics interface which emphasizes the significance of relational constraints, and treats purely syntactic, purely semantic and purely pragmatic constraints as parochial special cases. By making use of this perspective, COT is able to follow integrated grammar formalisms such as HPSG (Pollard and Sag, 1994) and LFG (Bresnan, 1982), in that it integrates information from different components, and suggests that syntax, semantics and pragmatics are mutually constraining. However, the use of bidirectional optimization pushes the interplay of constraints from different grammar components in an entirely novel direction. The consequences of moving in this direction remain to be fully evaluated.

7.5. Concluding Remarks

I have presented a framework in which theories of anaphora resolution can be developed, and I have argued that the framework provides fertile ground for further theoretical exploration of this and related issues. I have discussed some such areas, but omitted others of equal importance, such as the significance of rhetorical relations in discourse structure.[45]

---

[44] Arguably AGREE is non-relational: this would depend on whether agreement is purely formal, or involves some reference to meaning. Under the assumption that AGREE constrains the resolution of definite descriptions, which in English are not syntactically marked for gender, AGREE is a relational constraint.

[45] Grosz and Sidner's work (Grosz and Sidner, 1986) shows how a sophisticated theory of discourse structure is relevant to anaphora resolution. Aspects of their proposal are computationally implemented by Lochbaum (1998), although she does not provide a full analysis of anaphora resolution. Comparison of COT with the wide ranging proposals of Hobbs (Hobbs, 1985; Hobbs et al., 1993) and Asher and Lascarides (Lascarides and Asher, 1993; Asher and Lascarides, 1994) would also be valuable, although beyond the scope of the current work. The use of non-monotonic inference in these theories is reminiscent of the non-monotonicity inherent to OT. Connections with COT can probably best be seen via the intermediary of (de Hoop and de Swart, 1998), which makes a comparable use of discourse structure and temporal relations. Also in this regard, note that observations of Kehler (Kehler, 1997; Kehler, 1993) probably necessitate some introduction of rhetorical relations into the COT model.

It is obvious that such theoretical development must be accompanied by rigorous empirical work. Having motivated a set of constraints, we might concentrate research on determining the proper ordering of those constraints in a given language. For example, it would be natural to ask which ordering best captures the anaphoric relationships present in elicited text in a production experiment, or which best captures the anaphoric relationships in a tagged corpus. Indeed, given suitably tagged data sets it would be possible to induce an ordering automatically using techniques from machine learning (Tesar and Smolensky, 1998; Boersma and Hayes, 2001). In this way we could obtain an objective evaluation of the extent to which different theoretical ideas concerning discourse function significantly determine discourse form.

It is my hope that by enabling uniform description of both Centering and variants, the framework developed here will facilitate the empirical research that remains to be done.

Appendix

A. Equivalence of COT and BFP

It will now be shown that BFP and the version of COT from section 3 are extensionally equivalent in terms of core predictions about anaphora resolution.

The proof proceeds by dividing resolutions into three mutually exclusive cases:

1. There are syntactic violations and any combination of other constraint violations.

2. There is no syntactic violation, but there is a Rule 1 violation and any combination of other constraint violations.

3. There is no syntactic or Rule 1 violation, but there is some combination of other constraint violations.

The proof proceeds by showing that as regards resolutions in each of these three classes, COT and BFP make identical predictions.

In what follows, I use the term *purely anaphoric resolutions* to mean interpretations of a sentence in which all definites are interpreted as anaphoric on elements explicit in the previous sentence and such that no new information is conveyed by the definite. I also assume that there is no relevant restriction on available candidate interpretations.

PROPOSITION 1. *Purely anaphoric resolutions breaking syntactic constraints are never* COT *optimal, and never correspond to preferred* BFP *transitions.*

   Suppose that a purely anaphoric resolution R conflicts with at least one of AGREE or DISJOINT in COT. In such a case there are guaranteed to be alternative candidate interpretations which satisfy both of these constraints. To see this, observe that by breaking FAM-DEF repeatedly in some other candidate interpretation, we can ensure that no definite NPs are interpreted as anaphoric and that none corefer. AGREE and DISJOINT are then satisfied trivially.

   In BFP, interpretations failing agreement constraints are not produced in the *construction* phase of the algorithm, and interpretations failing syntactic coreference constraints are removed in the *filter* stage. In either case, such interpretations are ruled out before transition-based preferences between interpretations are even considered. So R will also not be the interpretation predicted by BFP. The reverse argument, that if syntactic violations rule a candidate out in BFP, it will also not be the optimal COT candidate, is similar.

PROPOSITION 2. *Fully anaphoric resolutions which violate Rule 1 are never* COT *optimal, and never correspond to preferred* BFP *transitions.*

   To violate Rule 1, there must be pronouns present. Consider a fully anaphoric resolution R of a sentence with pronouns, and suppose R violates Rule 1. Some non-pronominal element N must refer to a more salient entity than any pronominal element does. Clearly R also violates PRO-TOP. R cannot be COT optimal because there must be an alternative candidate R' in which all pronouns are resolved as in R, but in which no non-pronominal elements are interpreted anaphorically. Although R' will produce one or more violations of FAM-DEF, it will satisfy PRO-TOP, and can be chosen so as to satisfy the higher ranking syntactic constraints. The lower ranking of FAM-DEF than PRO-TOP guarantees that R' is a better candidate than R, so R is not COToptimal. It is also clear that R cannot be a preferred interpretation in BFP since candidates violating Rule 1 are filtered out from consideration in the second stage of the algorithm.

PROPOSITION 3. *Suppose two fully anaphoric resolutions A and B of a sentence satisfy syntactic constraints and Rule 1. If* COT *ranks candidate A above candidate B then* BFP *ranks candidate A above candidate B and* vice versa*.*

Suppose the sentence has pronouns. By assumption, A and B are fully anaphoric, so there can be no clash with FAM-DEF. Since Rule 1 is satisfied by A and B and there are pronouns, PRO-TOP is also satisfied by A and B. Given that syntactic constraints are also satisfied, it follows that violations of COHERE and ALIGN will determine which of A and B is preferred in COT, if either. Suppose there had been no pronouns. In this case syntactic constraints and FAM-DEF are sataisfied by both candidates, but PRO-TOP is violated by both. Again, violations of COHERE and ALIGN will determine which of A and B is preferred in COT.

In BFP, the fact that syntactic constraints are met and Rule 1 is satisfied means that both A and B will be subject to transition classification (regardless of whether pronouns are present). Thus the standard transition preferences will determine which of A and B is preferred in BFP, if either.

We have now established that in COT the choice between A and B will depend on COHERE and ALIGN, and in BFP he choice will depend on the transition preference ranking. The table in 19 lists all possible combinations of clashes with COHERE and ALIGN such that one candidate would be COT-preferred to another. For each of these, the corresponding BFP transition is uniquely determined, and in each case the top transition out-ranks the lower transition. This demonstrates the proposition from left to right. The reverse direction is similarly simple. The BFP column lists all the transition pairs such that one out-ranks the other. For each of these the pattern of COT clashes with COHERE and ALIGN is uniquely determined, and in each case the top one would be the COT preferred candidate. This demonstrates the proposition from right to left.

In combination with the first two parts of the proof, a more general claim follows, given earlier as (20):

PROPOSITION 4. *Given a sentence in which the only definite expressions are proper nouns and pronouns, if either* COT *(with the rankings in section 3) or* BFP *uniquely predicts an interpretation involving fully anaphoric interpretation of all definites, then both do, and in this case they resolve anaphors identically.*

References

Abbott, B.: 2001, 'Definiteness and Indefiniteness'. To appear: Laurence R. Horn and Gregory Ward (eds.), *The Handbook of Pragmatics*.
Aissen, J.: 1992, 'Topic and Focus in Mayan'. *Language* **68**(1), 43–80.

Aissen, J.: 1999, 'Markedness and Subject Choice in Optimality Theory'. *Natural Language and Linguistic Theory* **17**, 673–711.

Arnold, J.: 1998, 'Reference Form and Discourse Patterns'. Ph.D. thesis, Stanford University.

Aronoff, M.: 1976, *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Asher, N. and A. Lascarides: 1994, 'Intentions and Information in Discourse'. In: J. Pustejovsky (ed.): *Proceedings of the Thirty-Second Meeting of the Association for Computational Linguistics*. San Francisco, pp. 35–41, Morgan Kaufmann.

Beaver, D.: 1999a, 'The Logic of Anaphora Resolution'. In: P. Dekker (ed.): *Proceedings of the 12th Amsterdam Colloquium*. University of Amsterdam.

Beaver, D.: 1999b, 'Pragmatics (to a first approximation)'. In: J. Gerbrandy, M. Marx, M. de Rijke, and Y. Venema (eds.): *JFAK — Essays Dedicated to Johan van Benthem on the Occasion of his 50th Birthday*. Vossiuspers, Amsterdam University Press.

Beaver, D.: 2000, 'Pragmatics, and that's an order'. In: D. Barker-Plummer, D. Beaver, J. van Benthem, and P. S. di Luzio (eds.): *Symbols, Logic and Computation*. CSLI Press.

Blutner, R.: 2000, 'Some Aspects of Optimality in Natural Language Interpretation'. *Journal of Semantics* **17**(3), 189–216.

Blutner, R. and G. Jäger: 1999, 'Competition and interpretation: The German adverbs of repetition'. available at http://www2.hu-berlin.de/asg/blutner/.

Boersma, P. and B. Hayes: 2001, 'Empirical tests of the Gradual Learning Algorithm'. *Linguistic Inquiry* **32**, 45–86.

Bolinger, D.: 1977, *Meaning and Form*. New York: Longman.

Brennan, S.: 1995, 'Centering Attention in Discourse'. *Langauge and Cognitive Processes* **10**, 137–167.

Brennan, S., M. Friedman, and C. Pollard: 1987, 'A Centering Approach to Pronouns'. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Cambridge, Mass., Association for Computational Linguistics.

Bresnan, J. (ed.): 1982, *The Mental Representation of Grammatical Relations*. Cambridge, MA: The MIT Press.

Bresnan, J.: 1999, 'The Emergence of the Unmarked Pronoun'. In: G. Legendre, S. Vikner, and J. Grimshaw (eds.): *Optimality-theoretic Syntax*. MIT Press.

Brown, G. and G. Yule: 1983, *Discourse Analysis*. Cambridge University Press.

Buring, D.: 1999, 'On D-Trees, Beans, and B-Accents'. MS., UCSC.

Cahn, J.: 1995, 'The Effect of Pitch Accenting on Pronoun Referent'. In: *Proceedings of the 33rd International Joint Conference in Artificial Intelligence (Student Session)*. pp. 290–2, Association for Computational Linguistics.

Christopherson, P.: 1939, *The artocles: A study of their theory and use in English*. Copenhagen: Munksgaard.

de Hoop, H.: 2000, 'Optional Scrambling and Interpretation'. In: H. Bennis, M. Everaert, and E. Reuland (eds.): *Interface Strategies*. Amsterdam: KNAW, pp. 153–168.

de Hoop, H. and H. de Swart: 1998, 'Temporal adjunct clauses in Optimality Theory'. OTS Utrecht.

Dekker, P. and R. van Rooy: 2000, 'Optimality Theory and Game Theory: Some Parallels'. *Journal of Semantics* **17**(3), 217–242.

Eckert, M. and M. Strube: 2000, 'Dialogue Acts, Synchronising Units and Anaphora Resolution'. *Journal of Semantics.* To appear.

Givón, T.: 1983, 'Topic continuity in spoken discourse'. In: T. Givón (ed.): *Topic Continuity in Discourse: Quantified Cross-Linguistic Studies.* Amsterdam: John Benjamins, pp. 347–363.

Gordon, P. and D. Chan: 1995, 'Pronouns, Passives and Discourse Coherence'. *Journal of Memory and Language* **34**, 216–231.

Gordon, P., B. Grosz, and L. Gilliom: 1993, 'Pronouns, Names, and the Centering of Attention in Discourse'. *Cognitive Science* **17**(3), 311–347.

Gordon, P. and R. Hendrick: 1997, 'Intuitive Knowledge of Linguistic Co-reference'. *Cognition* **62**, 325–370.

Grimshaw, J.: 1997, 'Projection, Heads and Optimality'. *Linguistic Inquiry* **28**(3).

Groenendijk, J. and M. Stokhof: 1991, 'Dynamic Predicate Logic'. *Linguistics and Philosophy* **14**, 39–100.

Groenendijk, J., M. Stokhof, and F. Veltman: 1995, 'Coreference and Modality'. In: S. Lappin (ed.): *The Handbook of Contemporary Semantic Theory.* Oxford: Blackwell.

Grosz, B.: 1977, 'The representation and use of focus in a system for understanding dialogue'. In: *Proceedings of the Fifth International Joint Conference on Artificial Intelligence.* Cambridge, MA.

Grosz, B., A. Joshi, and S. Weinstein: 1983, 'Providing a Unified Account of Definite Noun Phrases in Discourse'. In: *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.* Cambridge, Mass., pp. 44–49, Association for Computational Linguistics.

Grosz, B., A. Joshi, and S. Weinstein: 1995, 'Centering: A Framework for Modeling the Local Coherence of Discourse'. *Computational Linguistics* **21**(2), 203–226.

Grosz, B. and C. Sidner: 1986, 'Attention, Intentions, and the Structure of Discourse'. *Computational Linguistics* **12**, 175–204.

Grosz, B. and C. Sidner: 1998, 'Lost Intuitions and Forgotten Intentions'. In: M. Walker, A. Joshi, and E. Prince (eds.): *Centering Theory in Discourse.* Oxford: Clarendon Press, pp. 89–112.

Gundel, J.: 1998, 'Centering Theory and the Givenness Hierarchy: Towards a Synthesis'. In: M. Walker, A. Joshi, and E. Prince (eds.): *Centering Theory in Discourse.* Oxford: Clarendon Press, pp. 359–400.

Gundel, J., N. Hedberg, and R. Zacharski: 1983, 'Cognitive Status and the Form of Referring Expressions in Discourse'. *Language* **69**, 274–307.

Heim, I.: 1982, 'On the semantics of Definite and Indefinite Noun Phrases'. Ph.D. thesis, Umass. Amherst.

Hendriks, P. and H. de Hoop: 2001, 'Optimality Theoretic Semantics'. *Linguistics and Philosophy* **24**(1), 1–32.

Hobbs, J.: 1985, 'The coherence and structure of discourse Technical Report CSLI-85-37'.

Hobbs, J., M. Stickel, D. Appelt, and P. Martin: 1993, 'Interpretation as abduction'. *Artificial Intelligence* **63**(1-2), 69–142.

Horn, L.: 1984, 'Toward a New Taxonomy for Pragmatic Inference: Q-Based and R-Based Implicature'. In: D. Schiffrin (ed.): *Meaning, Form, and Use in Context: Linguistic Applications.* Washington, DC: Georgetown University Press, pp. 11–42.

Hudson-D'Zmura, S.: 1989, 'Centering: A Framework for Modelling the Local Coherence of Discourse'. Ph.D. thesis, University of Rochester.

Joshi, A. and S. Kuhn: 1979, 'Centered Logic: The role of entity centered sentence representation in natural language inferencing'. In: *Proceedings of the 6th International Joint Conference in Artificial Intelligence*. Tokyo, pp. 435–9.

Joshi, A. and S. Weinstein: 1981, 'Control of Inference: Role of Some Aspects of Discourse Structure — Centering'. In: *Proceedings of the 7th International Joint Conference in Artificial Intelligence*. Vancouver, pp. 385–7, Association for Computational Linguistics.

Kameyama, M.: 1998, 'Intrasentential Centering: A Case Study'. In: M. Walker, A. Joshi, and E. Prince (eds.): *Centering Theory in Discourse*. Oxford: Clarendon Press, pp. 89–112.

Kameyama, M.: 1999, 'Stressed and unstressed pronouns: complimentary preferences'. In: P. Bosch and R. van der Sandt (eds.): *Focus: Linguistic, Cognitive and Computational Perspectives*. Cambridge University Press.

Kamp, H. and U. Reyle: 1993, *From Discourse to Logic*. Kluwer.

Katz, J.: 1980, 'Chomsky on Meaning'. *Language* **56**, 1–42.

Keenan, E.: 1976, 'Towards a universal definition of 'subject''. In: C. N. Li (ed.): *Subject and Topic*. New York: Academic Press, pp. 303–333.

Kehler, A.: 1993, 'Intrasentential Constraints on Intersentential Anaphora in Centering Theory'. Workshop on Centering Theory in Naturally Occurring Discourse, University of Pennsylvania.

Kehler, A.: 1997, 'Current Theories of Centering for Pronoun Interpretation: A Critical Evaluation'. *Computational Linguistics* **23**(3).

Kibble, R.: 1999, 'Cb or not Cb? Centering Theory applied to NLG'. Technical Report ITRI-99-17, University of Brighton. Presented at ACL Workshop on Discourse and Reference Structure, University of Maryland, June 1999.

Kiparsky, P.: 1983, 'Word-formation and the lexicon'. In: F. Ingeman (ed.): *Proceedings of the 1982 Mid-America Linguistic Conference*.

Kuno, S.: 1973, *The Structure of Japanese Language*. Cambridge, Mass.: MIT Press.

Kuno, S.: 1987, *Functional Syntax*. Chicago University Press.

Lambrecht, K.: 1994, *Information Structure and Sentence Form*. Cambridge University Press.

Langendoen, D. T.: 2000, 'An Optimality Theoretic account of the scope of operators'. University of Arizona.

Lascarides, A. and N. Asher: 1993, 'Temporal Interpretation, Discourse Relations and Common Sense Entailment'. *Linguistics and Philosophy* **16**(5), 437–493.

Lewis, D.: 1979, 'Scorekeeping in a Language Game'. *Journal of Philosophical Logic* **8**, 339–359.

Lochbaum, K. E.: 1998, 'A Collaborative Planning Model of Intentional Structure'. *Computational Linguistics* **24**(4), 525–572.

McCawley, J.: 1978, 'Conversational implicature and the lexicon'. In: P. Cole (ed.): *Syntax and Semantics 9: Pragmatics*. New York: Academic Press, pp. 245–259.

McNally, L.: 1998, 'On Recent Formal Analyses of Topic'. In: *The Tbilisi Symposium on Language, Logic, and Computation: Selected Papers*. Stanford, CA: CSLI Publications, pp. 147–160.

Nakatani, C.: 1997, 'The Computational Processing of Intonational Prominence: A Functional Prosody Perspective'. Ph.D. thesis, Harvard University. Available as CRCT technical report TR-15-97.

Pierrehumbert, J. and J. Hirschberg: 1990, 'The meaning of intonational contours in interpretation of discourse'. In: P. Cohen, J. Morgan, and M. Pollack (eds.): *Intentions in Communication*. Cambridge, Massachusetts: MIT Press.

Poesio, M., H. Cheng, R. Henschel, J. Hitzeman, R. Kibble, and R. Stevenson: 2000, 'Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation using Text from Application-Oriented Domains'. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Pollard, C. and I. Sag: 1994, *Head-Driven Phrase Structure Grammar.* University of Chicago Press and Stanford: CSLI Publications.

Portner, P. and K. Yabushita: 1998, 'The Semantics and Pragmatics of Topic Phrases'. *Linguistics and Philosophy* **21**(2), 117–157.

Prince, A. and P. Smolensky: 1993, 'Optimality Theory: Constraint Interaction in Generative Grammar. Technical Report 2'. Technical report, Rutgers University Center for Cognitive Science.

Prince, E.: 1981, 'Toward a taxonomy of given-new information'. In: P. Cole (ed.): *Radical Pragmatics.* New York: Academic Press, pp. 223–256.

Reinhart, T.: 1982, 'Pragmatics and Linguistics: An Analysis of Sentence Topics'. *Philosophica* **27**, 53–94.

Roberts, C.: 1998, 'The Place of Centering in a General Theory of Anaphora Resolution'. In: M. Walker, A. Joshi, and E. Prince (eds.): *Centering Theory in Discourse.* Oxford: Clarendon Press, pp. 359–400.

Schwarzschild, R.: 1999, 'Givenness, AvoidF and other Constraints on the Placement of Accent'. *Natural Language Semantics* **7**(2), 141–177.

Sells, P.: 1999, 'Form and Function in the Typology of Grammatical Voice Systems'. In: *Optimality-Theoretic Syntax.* Cambridge: MIT Press.

Sells, P.: 2000, 'Alignment Constraints in Swedish Clausal Syntax'. ms., University of Stanford.

Shi, D.: 2000, 'Topic and topic-comment in Mandarin Chinese'. *Language* **69**, 274–307.

Sidner, C. L.: 1983, 'Focusing in the Comprehension of Definite Anaphora'. In: M. Brady and R. C. Berwick (eds.): *Computational Models of Discourse.* Cambridge, MA: MIT Press, pp. 267–330.

Smolensky, P.: 1998, 'Why Syntax is Different (but not really): Ineffability, Violability and Recoverability in Syntax and Phonology'. Handout from talk at the Stanford University workshop: Is Syntax Different?

Strube, M.: 1998, 'Never Look Back: An Alternative to Centering'. In: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics.* Montreal, pp. 1251–1257.

Strube, M. and U. Hahn: 1999, 'Functional Centering: Grounding Referential Coherence in Information Structure'. *Computational Linguistics* **25**(3), 309–344.

Tesar, B. and P. Smolensky: 1998, 'Learnability in Optimality Theory'. *Linguistic Inquiry* **29**, 229–268.

Thompson, S. A.: 1987, 'The passive in English: A discourse perspective'. In: R. Chan-non and L. Shockey (eds.): *In Honor of Ilse Lehiste.* Dordrecht: Foris, pp. 497–511.

Valduví, E. and M. Vilkuna: 1997, 'On Rheme and Kontrast'. In: P. Culicover and L. McNally (eds.): *The limits of syntax.* New York: Academic Press.

Vallduví, E.: 1990, 'The Informational Component'. Ph.D. thesis, UPenn. Philadelphia.

Vallduvi, E.: 1993, 'Information packaging: A survey'. Technical Report HCRC/RP-44, University of Edinburgh.

van der Does, J. and H. de Hoop: 1998, 'Type-Shifting and Scrambled Definites'. *Journal of Semantics* **15**, 393–416.

Walker, M., M. Iida, and S. Cote: 1994, 'Japanese Discourse and the Process of
    Centering'. *Computational Linguistics* **20/2**, 193–232.
Walker, M., A. Joshi, and E. Prince (eds.): 1998, *Centering Theory in Discourse*.
    Oxford: Clarendon Press.
Ward, G.: 1988, *The Semantics and Pragmatics of Preposing*. New York: Garland.
Zeevat, H.: 1999, 'Explaining Presupposition Triggers'.
Zeevat, H.: 2000, 'The assymmetry of optimality theoretic syntax and semantics'.
    Journal of Semantics **17**(3), 243–262.