Current Theories of Centering for Pronoun Interpretation: A Critical Evaluation

Andrew Kehler *
SRI International

We review the fundamental concepts of centering theory, and discuss some facets of the pronoun interpretation problem that motivate a centering-style analysis. We then demonstrate some problems with a popular centering-based approach with respect to these motivations.

1. Introduction

A central claim of centering theory (Grosz, Joshi, and Weinstein, 1995, henceforth GJW) is that certain entities mentioned in an utterance are more central than others, and that this property imposes constraints on a speaker's use of different types of expressions to refer to them. To articulate some of these constraints, they define several fundamental centering concepts and propose rules based on them that should be followed by a speaker in producing coherent discourse. This work has led to several analyses employing centering theory and extensions of it, particularly in the area of pronoun interpretation (Kameyama, 1986; Brennan, Friedman, and Pollard, 1987; Di Eugenio, 1990; Walker, Iida, and Cote, 1994; Di Eugenio, 1996; Strube and Hahn, 1996, inter alia; see also citations within GJW, forthcoming papers in Walker, Joshi, and Prince (in press), and psycholinguistic studies described in Hudson-D'Zmura (1989), Gordon, Grosz, and Gilliom (1993), and Brennan (1995)).[1] In this squib, we discuss some facets of the pronoun interpretation problem that motivate a centering-style analysis, and demonstrate some problems with a popular centering-based approach with respect to these motivations.

2. Overview of Centering

Centering theory is motivated by two related facts about language that are not explained by purely content-based models of reference and coherence (cf. Hobbs (1979)): (1) that the coherence of a discourse does not depend only on semantic content but also on the type of referring expressions used, and (2) the existence of garden path effects, in which pronouns appear to be resolved before adequate semantic information has become available:

> Pronouns and definite descriptions are not equivalent with respect to their effect on coherence. We conjecture that this is so because they engender different inferences on the part of a hearer or reader. In the most pronounced cases, the wrong choice will mislead a hearer and force

* Artificial Intelligence Center, 333 Ravenswood Avenue, Menlo Park, CA 94025; kehler@ai.sri.com.

1 A draft of GJW, which revised and expanded ideas presented in Grosz, Joshi, and Weinstein (1983), was circulated as far back as 1986. Therefore some of the works described here as extending the work contained therein are dated prior to the published version.

> backtracking to a correct interpretation. (Grosz, Joshi, and Weinstein, 1995, pg. 207)

GJW exemplify the first of these motivations with passages (1) and (2). Passage (1) is presumed to be in a longer segment that is currently centered on John.

(1) a. He has been acting quite odd. (*He*=John)

b. He called up Mike yesterday.

c. John wanted to meet him quite urgently.

The third sentence in this passage is quite odd, presumably because the more central element (John) is not referred to with a pronoun whereas the less central element (Mike) is. This passage can be compared to the similar passage in (2).

(2) a. He has been acting quite odd. (*He*=John)

b. He called up Mike yesterday.

c. He wanted to meet him quite urgently.

Although the propositional content expressed by these two passages is the same (the only difference being the expression used to refer to John in the subject of the third sentence), passage (2) is not jarring in the way that (1) is.

GJW exemplify the second of these motivations with passage (3).

(3) a. Terry really goofs sometimes.

b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.

c. He wanted Tony to join him on a sailing expedition.

d. He called him at 6AM.

e. He was sick and furious at being woken up so early.

Sentence (3e) causes the hearer to be misled: whereas common sense considerations indicate that the intended referent for *He* is Tony, hearers tend to initially assign Terry as its referent. Such examples suggest that more is involved in pronoun interpretation than simply reasoning about semantic plausibility. In particular, they suggest that hearers assign referents to pronouns before interpreting the remainder of the sentence.

Details of Centering. In GJW's centering theory, each utterance U_n in a discourse has exactly one backward-looking center (denoted $C_b(U_n)$) and a partially ordered set of forward-looking centers (denoted $C_f(U_n)$). Roughly speaking, $C_f(U_n)$ contains all entities referred to in U_n; among these is $C_b(U_n)$. Following Brennan, Friedman, and Pollard (1987), we refer to the highest-ranked forward-looking center as $C_p(U_n)$.[2] $C_b(U_{n+1})$ is by definition the most highly ranked element of $C_f(U_n)$ realized in U_{n+1}. Three intersentential relationships between a pair of utterances U_n and U_{n+1} are defined:

2 The issues pertaining to how the ordering of entities in $C_f(U_n)$ is determined have not been completely resolved. For the examples discussed in this paper, we can use the hierarchy of grammatical relations given by Brennan, Friedman, and Pollard (1987), in which the grammatical subject is ranked above all other grammatical relations (object, object2, and so forth).

1. Center Continuation: $C_b(U_{n+1}) = C_b(U_n) = C_p(U_{n+1})$. In this case $C_b(U_{n+1})$ is the most likely candidate for $C_b(U_{n+2})$.
2. Center Retaining: $C_b(U_{n+1}) = C_b(U_n)$, but $C_b(U_{n+1}) \neq C_p(U_{n+1})$. In this case $C_b(U_{n+1})$ is not the most likely candidate for $C_b(U_{n+2})$.
3. Center Shifting: $C_b(U_{n+1}) \neq C_b(U_n)$.

The following rules are proposed in GJW:

> Rule 1: If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.
>
> Rule 2: Sequences of continuations are preferred over sequences of retaining; and sequences of retaining are to be preferred over sequences of shifting.

The use of Rule 1 is illustrated by the oddness of passage (1) as compared to passage (2), because in (1c) the C_b (John) is not pronominalized whereas a non-C_b (Mike) is. The examples GJW give to illustrate Rule 2 are shown in passages (4) and (5).

(4) 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 that he could finally buy a piano.
d. He arrived just as the store was closing for the day.

(5) a. John went to his favorite music store to buy a piano.
b. It was a store John had frequented for many years.
c. He was excited that he could finally buy a piano.
d. It was closing just as John arrived.

Like passages (1) and (2), passages (4) and (5) express the same propositional content, yet they are not equally coherent. Whereas passage (4) consists of a sequence of Continue relations centered on John, passage (5) is marked by movements between Continuing and Retaining, which gives the effect that the passage flips back-and-forth between being about John and being about his favorite music store.

Rule 1 is presented as a constraint on center realization, and Rule 2 as a constraint on center movement. As formulated, the predictions these rules make about the preferred referents of pronouns are fairly limited.[3] For instance, Rule 1 makes no predictions about the preferred referents of the pronouns in sentence (3d), nor does it predict the garden path effect in sentence (3e); in each case the rule is satisfied assuming either possible assignment of referents to the pronouns.[4]

3 GJW do not make any specific proposals for using Rules 1 and 2 for pronoun interpretation. In Section 3, we discuss a particular utilization of these rules for pronoun interpretation proposed by Brennan, Friedman, and Pollard (1987). An apparently popular misconception attributes this utilization to GJW, however neither the draft nor final versions of GJW put forth such a proposal. See also GJW (1995, pg. 215, footnote 16).

4 A case in which Rule 1 does make a prediction is given in example (6); assigning Sam as the referent of *he* causes a violation whereas assigning John does not.

(6) a. John introduced Bill to Sam.
b. He seemed to like Bill.

I thank an anonymous reviewer for bringing this example to my attention.

3. The BFP Algorithm

Brennan, Friedman, and Pollard (1987, henceforth BFP) describe an algorithm for pronoun interpretation based on centering principles, which is also utilized in Walker, Iida, and Cote (1994, henceforth WIC). In addition to Rule 1, BFP utilize Rule 2 in making predictions for pronominal reference. They augment the transition hierarchy by replacing the Shift transition with two transitions, termed Smooth-Shift and Rough-Shift, which are differentiated on the basis of whether or not $C_b(U_{n+1})$ is also $C_p(U_{n+1})$.[5]

3a. Smooth-Shift: $C_b(U_{n+1}) = C_p(U_{n+1})$, $C_b(U_{n+1}) \neq C_b(U_n)$.

3b. Rough-Shift: $C_b(U_{n+1}) \neq C_p(U_{n+1})$, $C_b(U_{n+1}) \neq C_b(U_n)$.

They redefine Rule 2 as follows:

> Rule 2: Transition states are ordered. CONTINUE is preferred to RETAIN is preferred to SMOOTH-SHIFT is preferred to ROUGH-SHIFT.

The resulting transition definitions are summarized in Table 1.

	$C_b(U_{n+1}) = C_b(U_n)$ or unbound $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Table 1
Transitions in the BFP Algorithm

Given these definitions, their algorithm (as described in WIC) is defined as follows.

1. GENERATE possible C_b-C_f combinations
2. FILTER by constraints, e.g., contra-indexing, sortal predicates, centering rules and constraints
3. RANK by transition orderings

The pronominal referents that get assigned are those which yield the most preferred relation in Rule 2, assuming Rule 1 and other coreference constraints (gender, number, syntactic, semantic type of predicate arguments) are not violated. This strategy correctly predicts that *He* and *him* in sentence (3d) refer to Terry and Tony respectively, since this assignment results in a Continue relation whereas the Tony/Terry assignment results in a less-preferred Retain relation. Their rules also account for the oddness of sentence (3e), since assigning *he* to Tony results in a Smooth-Shift whereas assigning *he* to Terry results in a Continue. Therefore, the algorithm makes the correct predictions regarding example (3), one of the central motivating examples of centering theory.

Problems with the BFP Algorithm. The fact that the BFP algorithm predicts the garden path effect exhibited by sentence (3e) is particularly indicative that it embodies the motivations for centering theory. As we noted in Section 2, such effects distinguish centering-based approaches from purely content-based models of reference and coherence (Hobbs, 1979, inter alia). As Brennan (1995) explains:

5 The terms Smooth-Shift and Rough-Shift were introduced in WIC.

> While knowledge-based theories often succeed in resolving referring expressions in this manner [=using semantic information and world knowledge, without taking advantage of the kinds of syntactic constraints that centering uses], they do not model human discourse processing. An entirely knowledge-based algorithm would not reproduce an addressee's immediate tendency to interpret a pronoun as cospecifying the backward center, even when this results in an implausible interpretation. (Brennan, 1995, pg. 145)

However, other examples demonstrate that the BFP algorithm also cannot model an addressee's immediate tendency to interpret a pronoun, and therefore cannot properly account for the pronoun interpretation preferences that result from such tendencies.

To illustrate, we consider a modification to passage (3), shown in passage (7), with three possible follow-ons ($7e_1$-e_3).

(7) a. Terry really gets angry sometimes.

b. Yesterday was a beautiful day and he was excited about trying out his new sailboat.

c. He wanted Tony to join him on a sailing expedition, and left him a message on his answering machine. [C_b=C_p=Terry]

d. Tony called him at 6AM the next morning. [C_b=Terry, C_p=Tony]

e_1. He was furious for being woken up so early.

e_2. He was furious with him for being woken up so early.

e_3. He was furious with Tony for being woken up so early.

Sentence (7d) constitutes a Retain, in which $C_p(U_{7d})$ is Tony and $C_b(U_{7d})$ is Terry. Retains often result in an ambiguity based on whether a subsequent subject pronoun refers to $C_b(U_n)$ (resulting in a Continue) or to $C_p(U_n)$ (resulting in a Smooth-Shift). While the subject pronouns in follow-ons ($7e_1$-e_3) may all display this ambiguity to a certain degree, the preferences associated with them appear to be consistent among the three variants.[6] That is, the initial preference for the subject pronominal *He* in sentence ($7e_1$) does not appear to be affected by the subsequent inclusion of the phrases *with him* in variant ($7e_2$) and *with Tony* in variant ($7e_3$). This accords with the observation that hearers have an immediate tendency to resolve subject pronouns based on the existing discourse state, before the entire sentence is interpreted.

However, the ways in which these follow-ons are analyzed within the BFP algorithm differ radically, as summarized in Table 2. In follow-on ($7e_1$), assigning *He*=Terry results in a Continue whereas assigning *He*=Tony results in a Smooth-Shift, and so Terry is preferred. In follow-on ($7e_2$), assigning *He*=Terry results in a Rough-Shift whereas assigning *He*=Tony again results in a Smooth-Shift, and so Tony is preferred. The reason for this difference is attributable solely to the fact that the pronoun *him* occurs in ($7e_2$):

6 The author and several informants prefer the subject pronoun to refer to Tony initially, causing a garden path effect in each case. Aside from this, there may be a subtle processing difference between these sentences in that any garden path in sentence ($7e_3$) may be resolved earlier than in ($7e_1$) and ($7e_2$), specifically, at the point at which *Tony* is reached. This is a result of the fact that syntactic constraints on coreference can be used to eliminate the possibility of *He* referring to Tony at that time, whereas in the other cases it is semantic information that comes later in the sentence that eliminates Tony as a referent.

Sentence	Subject Referent	$C_b(U_{7e_i})$	$C_p(U_{7e_i})$	Result	Preference
7e$_1$	Terry	Terry	Terry	Continue	Terry
	Tony	Tony	Tony	Smooth-Shift	
7e$_2$	Terry	Tony	Terry	Rough-Shift	Tony
	Tony	Tony	Tony	Smooth-Shift	
7e$_3$	Terry	Tony	Terry	(#) Rule 1 Violation (Rough Shift)	??
	Tony	Tony	Tony	(*) Condition C Violation	

Table 2
Centering Analysis of Sentences (7e$_1$-e$_3$)

because there are two non-coreferring pronouns in (7e$_2$), one must refer to Tony, and because Tony is $C_p(U_{7d})$, by definition Tony is $C_b(U_{7e_2})$ instead of Terry. Finally, in sentence (7e$_3$), the assignment of *He*=Terry results in a Rule 1 violation – the C_b Tony is not pronominalized whereas Terry is – putting it in the company of highly awkward examples such as passage (1). If we ignore this violation, the resulting transition is again a Rough-Shift, the lowest-ranked relation. (The assignment of *He*=Tony is ruled out by a syntactic constraint violation.)

These varied results are inconsistent with the aforementioned facts concerning these passages in both empirical and theoretical respects. Empirically, the results are counter to the more consistent preferences associated with the subject pronouns in each case. Theoretically, such consistency is just what one would expect given a hearer's immediate tendency to resolve subject pronouns based on the existing discourse state. In either regard, it is unclear why the inclusion of the phrases *with him* in variant (7e$_2$) and *with Tony* in variant (7e$_3$) should lead to such varied predictions for the subject pronoun. In fact, the example illustrates a general property of the BFP algorithm: that the preferred assignment for a pronoun in such examples, even in subject position, cannot be determined until the entire sentence has been processed. This property results from the fact that determining the transition type between a pair of utterances U_n and U_{n+1} requires the identification of $C_b(U_{n+1})$, and a noun phrase (pronominal or not) can occur *at any point in the utterance* that will alter the assignment of $C_b(U_{n+1})$. This is what occurs in the analysis of passage (7): whereas the C_b of sentence (7e$_1$) is Terry assuming *He* refers to Terry, the occurrence of *him* later in the sentence in (7e$_2$) and similarly *Tony* in (7e$_3$) cause the C_b to be Tony, thus changing the bindings that constitute the various transition possibilities, and in this case, the predicted preferred referents. To be clear, this is not an issue regarding the efficiency nor the cognitive reality of BFP's particular algorithm; in fact neither BFP nor WIC make any claims to these effects. The problem lies more generally in their proposal to utilize Rule 2 along with the definition of $C_b(U_{n+1})$ to interpret pronouns – *any* algorithm incorporating this proposal will have to process an entire sentence before determining the preferred referents of pronouns; no reordering of processing within the BFP algorithm can alter this fact. The need to process an entire sentence to recover pronoun assignments, however, is one that GJW and Brennan (1995) argue against in motivating centering over purely content-based models of reference and coherence. That is, this very property renders such an approach incapable of modeling the preferences associated with an addressee's immediate tendency to interpret pronouns, as example (7) demonstrates.[7]

Preferences and Other Intersentential Relationships. The motivations for centering cited by GJW and Brennan (1995) reflect the intuition that salience plays a central role in pronoun interpretation. What remains at issue is the manner in which salience is utilized by the pronoun interpreter. In the previous section we argued that BFP's use of Rule 2 along with the transition definitions and definition of C_b does not provide the correct utilization. In fact, the only aspects of U_n and U_{n+1} utilized by the BFP algorithm are the identities of $C_b(U_n)$, $C_p(U_n)$, $C_b(U_{n+1})$, and $C_p(U_{n+1})$, as well as the types of expressions used to refer to them. Here, we argue that this is also insufficient.

There is a well-known contrast between passages that are coherent by virtue of being a *narration*, as is the case for sentence (8c) and follow-on (8d), versus those coherent by virtue of *parallelism*, as is the case for sentence (8c) and follow-on (8d′).

(8) a. The three candidates had a debate today.

b. Bob Dole began by bashing Bill Clinton.

c. He criticized him on his opposition to tobacco.

d. Then Ross Perot reminded him that most Americans are also anti-tobacco.

d′. Then Ross Perot slammed him on his tax policies.

The preferred referent for the pronoun in example (8d) is Bob Dole, whereas the preferred referent for the pronoun in example (8d′) is Bill Clinton. However, each passage shares sentences (8a-c), and therefore $C_p(U_{8c})$ and $C_b(U_{8c})$ are the same for each follow-on. Furthermore, each follow-on contains a new subject (Ross Perot, who will be the new C_p) and an object pronoun (the referent of which will be the new C_b). Therefore, because the relevant C_b and C_p relations are the same, a BFP-style approach cannot distinguish between these cases.[8] These examples show that pronominal reference preferences are affected by additional types of intersentential relationships that may be identifiable at the time a pronoun is encountered; proposals along these lines include preference-ranking schemes (e.g., Kameyama (1996)) and systems in which salience and the process of determining coherence relations interact (e.g., Kehler (1995)).

7 In order to model this tendency in the BFP algorithm, one might consider a strategy in which *provisional* referents are assigned to pronouns while proceeding left-to-right in the current utterance. Under such a strategy one could assume that $C_b(U_{n+1})$ is computed incrementally using the assumption that no additional elements will appear in U_{n+1} that are more highly ranked in $C_f(U_n)$. Then, garden paths would be predicted when this assumption does not hold and the assignment of $C_b(U_{n+1})$ must be changed, in addition to those caused by semantic influences such as in sentence (3e).

Again, however, this strategy would treat follow-ons ($7e_1$) and ($7e_2$) quite differently. This strategy would predict no garden path effect for follow-on ($7e_1$), since it assigns Terry as the referent of *he* and sticks with it. On the other hand, ($7e_2$) should be much worse because two garden paths would be predicted: one for changing $C_b(U_{n+1})$ from Terry to Tony when the pronoun *him* is processed, and another for the semantic information subsequently preferring Terry. This difference does not appear to be reflected in the actual judgements for these two examples (in both cases we find a similar garden path effect), although experimental evidence would be required to confirm these judgements.

8 The BFP approach prefers Bob Dole as the referent for the pronoun in each case. Note that passage (8) with follow-on (8d′) contradicts BFP's (pg. 157) and WIC's (pg. 223) claim that constraints based on structural parallelism, such as Kameyama's (1986) property-sharing constraint, are epiphenomena of BFP's ordering of the C_f and preference for Continue interpretations, since such constraints predict that Bill Clinton is the referent of *him*, not Bob Dole. Note also that an appeal to semantic plausibility factors to alter the preferences for example (8d′) will not work, since it is at least as plausible that Perot would slam Dole on his tax policies as it is that he would slam Clinton.

Suri and McCoy (1994) also provide minimal pairs that are problematic for BFP, which their algorithm correctly handles. However, their algorithm also cannot distinguish between the above pair, preferring *Bill Clinton* in both cases.

4. Conclusions

The pronoun resolution preferences that result from an addressee's immediate tendency to interpret a pronoun motivate pursuing a centering-based approach. However, certain examples demonstrate that BFP's utilization of the centering rules does not model this tendency, which in turn limits the ability of their algorithm to account for the data. Furthermore, data has been presented that show that in addition to the salience factors utilized by BFP, additional types of intersentential relationships must be taken into account.

Acknowledgements

The author thanks Barbara Grosz, David Israel, Megumi Kameyama, Christine Nakatani, Gregory Ward, and four anonymous reviewers for helpful comments and discussions. This research was supported by National Science Foundation/Advanced Research Projects Agency Grant IRI-9314961 to SRI International and National Science Foundation Grant IRI-9404756 to Harvard University.

References

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive Processes*, 10:137–167.

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Meeting of the Association for Computational Linguistics*, pages 155–162.

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

Di Eugenio, Barbara. 1996. The discourse functions of italian subjects: a centering approach. In *Proceedings of the International Conference on Computational Linguistics (COLING-96)*.

Gordon, Peter C., Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17(3):311–347.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in English. In *Proceedings of the 21st Conference of the Association for Computational Linguistics (ACL-83)*, Cambridge, MA.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2).

Hobbs, Jerry. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Hudson-D'Zmura, Susan. 1989. *The Structure of Discourse and Anaphor Resolution: The discourse center and the roles of nouns and pronouns*. Ph.D. thesis, University of Rochester.

Kameyama, Megumi. 1986. A property-sharing constraint in centering. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, New York.

Kameyama, Megumi. 1996. Indefeasible semantics and defeasible pragmatics. In M. Kanazawa, C. Piñon, and H. de Swart, editors, *Quantifiers, Deduction, and Context*. CSLI, Stanford, CA, pages 111–138.

Kehler, Andrew. 1995. *Interpreting Cohesive Forms in the Context of Discourse Inference*. Ph.D. thesis, Harvard University.

Strube, Michael and Udo Hahn. 1996. Functional centering. In *Proceedings of the 34th Conference of the Association for Computational Linguistics (ACL-96)*, Santa Cruz, CA, June.

Suri, Linda Z. and Kathleen F. McCoy. 1994. RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics (Squibs and Discussions)*, 20(2):301–317.

Walker, Marilyn, Aravind Joshi, and Ellen Prince, editors. In press. *Centering in Discourse*. Oxford University Press.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2).