Intrasentential Constraints on Intersentential Anaphora in Centering Theory

Andrew Kehler
Harvard University
Aiken Computation Laboratory
33 Oxford Street
Cambridge, MA 02138
kehler@das.harvard.edu

February 26, 1993

1 Introduction

Centering has been proposed as a theory of how referring expressions interact with mechanisms used to maintain coherence in discourse [3, 4, 6]. The theory is primarily generation-oriented; its rules make their strongest predictions from the perspective of producing discourse. The theory makes less strong predictions on the *interpretation* side, for instance, in predicting the referents of pronouns.

In this abstract, I consider the interpretation side of Centering Theory, and propose rules that make stronger predictions for pronoun reference resolution. In Section 2, I provide background to Centering Theory. The relationship between Centering and agreement and syntactic coreference constraints is discussed in Section 3. In Section 4, I present some new data shedding light on the pronoun interpretation problem, and propose Centering and syntactic coreference rules that account for them. I conclude in Section 5.

2 Centering Theory

In the Centering Theory of Grosz et al. (henceforth GJW) [4], each utterance in a discourse has exactly one backward-looking center (denoted Cb) and a partially-ordered set of forward-looking centers $(Cf^1, ..., Cf^n)$. The notation Cb_n is used to refer to the Cb of sentence n, and Cf_n likewise to the Cf list of sentence n. Following [1], I refer to the highest-ranked forward-looking center Cf_n^1 as Cp_n. Roughly speaking, Cf_n contains all entities that can be referred to in sentence n; amongst this list is Cb_n. If Cb_{n+1} refers to anything, it must refer to a member of Cf_n, i.e., not members of preceding Cf lists. The rules specifying how entities are ranked on the Cf list are not fully developed, but factors that have been suggested to affect ranking include grammatical role, thematic role, and pronominalization. Three intersentential relationships between a pair of sentences n and $n+1$ are defined:

1. Center Shifting: $Cb_{n+1} \neq Cb_n$.
2. Center Retaining: $Cb_{n+1} = Cb_n$, but $Cb_n \neq Cp_n$. This may be seen as a preparation for a Shift.
3. Center Continuation: $Cb_{n+1} = Cb_n = Cp_n$.

The following rules are stipulated:

Rule 1: Cb_{n+1} = the highest ranked member of Cf_n that is realized in utterance *n+1*. If any other members of Cf_n are realized in utterance *n+1* as a pronoun, then Cb_{n+1} must be realized as a pronoun also.

Rule 2: Sequences of Continuations are preferred over sequences of Retentions which are in turn preferred over repeated Shifts.

The way in which multiple pronouns are assigned referents is left unspecified as long as Rule 1 is followed. GJW notes that in example (1), the reference in sentence (1e) causes the reader to be misled:

(1) 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 begin woken up so early.

Example (1) shows that there is more to pronoun determination than intersentential coherence determination alone (cf. [5, 11]). However, Rule 1 of Centering Theory doesn't predict with which entities in sentence (1c) the pronouns in sentence (1d) corefer, since both references in sentence (1d) are pronominalized. That is, even if "He" refers to *Tony* and "him" refers to *Terry*, then Rule 1 remains satisfied. Likewise in sentence (1e), Rule 1 doesn't make any prediction since there is only one pronominalized entity.

The algorithm of Brennan, Friedman, and Pollard (henceforth BFP) [1] utilizes Rule 2 in making predictions for pronominal reference. They augment the hierarchy with a fourth relation, termed *Shifting-1*:

- Center Shifting-1: $Cb_n = Cp_n$, $Cb_n \neq Cb_{n-1}$.

The notion of *Shifting* is hence redefined:

- Center Shifting: $Cb_n \neq Cp_n$, $Cb_n \neq Cb_{n-1}$.

Both of these are *Shifts* in the original rule; BFP split this category into two on the basis of whether or not the backward-looking center is the highest-ranked member of the Cf list. Rule 2 is then rewritten:

- Rule 2: Continuing is preferred over Retaining, which is preferred over Shifting-1, which is preferred over Shifting.

The assignment of pronominal referents is that which yields the most preferred relation in Rule 2 (assuming Rule 1 is followed). This correctly predicts the assignment of referents of the pronouns in sentence (1d), since this assignment results in a *Continue* relation while the incorrect assignment results in a less-preferred *Retain* relation. Their rules also account for the oddness of sentence (1e), since *he* coreferring with *Tony* constitutes a *Shift* relation whereas coreferring to *Terry* constitutes a *Continue* relation.

3 Constraints on Coreference

In my study, I seek to give an analysis for pronoun resolution that is not reliant on the more general rule for coherence given by Rule 2. In this section I note some constraints that must augment Centering Theory in its determination of pronominal referents. While these constraints are well-known and are generally employed in systems (e.g., [1]), the discussion will lay the groundwork for a more extensive proposal given in Section 4.

First, the extent to which Shifts cause additional processing load depends on the agreement features of the pronouns (i.e., gender, number, person). If example (1) is modified so as to make Terry female, then there is apparently little difficulty in interpretation despite the Shift in the final sentence:

(2) a. Terry really goofs sometimes.
b. Yesterday was a beautiful day and she was excited about trying out her new sailboat.
c. She wanted Tony to join her on a sailing expedition.
d. She called him at 6AM.
e. He was sick and furious at being woken up so early.

Unlike the garden path effect caused by sentence (1e), the identical (2e) is easily processable since the pronoun has only one possible referent.

Second, as is well-established [2, 7, 8, 10], intrasentential coreference constraints may rule out possible reference assignments. For example, in passage (3),

(3) No one spoke to Oscar yesterday at Ernie's party. Oscar felt sorry for him.

the pronoun *him* cannot refer to *Oscar*, because a reflexive would have to be used. Only the less sensical reading where *Oscar* felt sorry for *Ernie* is possible, even though the most salient object in the first sentence is *Oscar*.[1] As noted by Lasnik [8], no 'accidental' coreference can occur – apart from the fact that *him* cannot refer to the instance of *Oscar* in its own sentence, neither can it obtain its referent from the instance of *Oscar* in preceding discourse. Of course, a similar reference is allowable with a different entity in place of *Oscar*:

[1]That is, if the sentence started with *He*, the reader would at least initially be inclined to assume that Oscar is the referent.

(4) No one spoke to Bert$_i$ yesterday at Ernie's party. Oscar felt sorry for him$_i$.

In sum, the accessibility of referents in light of agreement and intrasentential coreference constraints are key in determining when an NP can be pronominalized. That is, it appears that as long as an NP is coreferent with the most highly-ranked member of the preceding Cf list *for which no constraints apply* (I henceforth refer to this Cf member as an entity's *preferred referent*), then that NP can be pronominalized (regardless of the relation in Rule 2). This observation does not yield a full pronominal reference resolution algorithm in itself, since there are cases (such as sentences with multiple pronouns) where a subset of the pronouns may not refer to their preferred referents as defined here (as in example (5) in the next section), but nevertheless are easily processable. In the following section, I claim that with a straightforward alteration of a commonly assumed classification of pronoun types, pronouns can be shown to have a strong bias for referring only to their preferred referents.

4 Additional Constraints on Coreference

The difference between example (5) and examples (6) and (7) is the central impetus for the theory that follows; I focus on the interpretation of the second (underlined) pronoun assuming that in each case the first pronoun refers to *Terry*:

(5) Terry talked to Sandy.
He said that <u>he</u> should quit his job.

(6) Terry has been getting harassed by Sandy at school.
Her father told <u>her</u> to stop it.

(7) Terry's parents have been socializing with Sandy's parents lately.
Yesterday, his mother went to the movies with <u>his</u> father.

In example (5), the second pronoun may refer to either *Terry* or *Sandy* without processing difficulty. In contrast, in examples (6) and (7) there is a strong bias for assigning the second pronoun to *Terry*; as a result these examples seem odd because context strongly prefers that *Sandy* be the referent.

The following two rules account for this data in a straightforward way:

(a) Intersententially-referring pronouns have a strong bias towards their *preferred referent*, that is, the most highest-ranked Cf entity for which reference is not blocked by syntactic coreference or agreement constraints.

(b) Pronouns of all cases have reflexive and non-reflexive forms (e.g., accusative = himself/him, nominative = he/he, genitive = his/his).

We use these two rules in conjunction with the following constraints, adapted from Binding Theory [2]:[2]

[2] Besides treating nominative and genitive pronouns as ambiguous between reflexive and non-reflexive forms, we also break from Binding Theory in not employing the notion of *minimal domains*. A discussion of the impact of

(c) Non-reflexive pronouns *cannot* corefer with a c-commanding NP.

(d) Reflexive forms *must* corefer with a c-commanding NP.

Let us now consider how this account makes the correct predictions for examples (5)-(7). In all of these examples, the first pronoun in not c-commanded by anything, so each can only be the non-reflexive form. Since there are no agreement or syntactic constraints that apply, each refers to the most accessible entity, *Terry*.[3] In example (5), the second pronoun is c-commanded by the first; therefore it is ambiguous between its reflexive and non-reflexive forms. The reflexive form is not constrained against referring to the most salient object *Terry* (in fact it must do so, since it must corefer with its c-commanding NP). The non-reflexive form *cannot* corefer with its c-commanding NP. Because there is a syntactic constraint against it referring to *Terry*, it must refer to the next most salient entity, *Sandy*. Therefore, the account correctly predicts that the second pronoun may refer to either entity, since the reflexive form has *Terry* as its preferred referent whereas the non-reflexive form has *Sandy* as its preferred referent.

For both examples (6) and (7), the second pronoun is not c-commanded by the first. Therefore, the reflexive form of the pronoun cannot exist there.[4] For the non-reflexive form, there are no syntactic conditions on coreference possibilities, so its preferred referent is the Cp, namely *Terry*. Therefore, there is no possible form of the pronoun that has *Sandy* as its preferred referent, so this theory accounts for the difficulty in assigning that referent in examples (6) and (7).

As these examples show, there is a strong preference for a pronoun to refer to the most accessible entity that it can; the ambiguity of the second pronoun in example (5) falls out from the categorization of pronouns into reflexive and non-reflexive forms. I note that examples (6) and (7) further show that there is more to determining coreference than computing coherence relations [5, 11]. In each of these examples, assigning the referent of the second pronoun to *Sandy* (unlike assigning it to *Terry*) results in perfectly well-formed and coherent text. Also, in general there is a strong bias for matrix-initial subjects to refer to the most accessible entity in the preceding Cf list (modulo agreement). This effect was apparent in example (1), and is captured in the BFP algorithm by preferring the *Continue* relation over others. This prediction naturally results from my rules: since a matrix-initial subject is not c-commanded by anything, only the non-reflexive form can exist, resulting in no syntactic constraints disallowing reference to its most accessible entity.[5] Furthermore, it is possible to use intonation to override a preferred referent. For example, stress may be used to indicate that the second *his* in example (8) refers to *Sandy*; we take the need for such stress to be evidence that the preferred referents do in fact exist:

---

these rules on intrasentential anaphora is beyond the scope of this abstract. However, in his theory of intrasentential anaphora, Kuno [7] proposes that genitive and accusative pronouns without the *-self* suffix are ambiguous between reflexive and non-reflexive forms, and also eschews the notion of minimal domain. Kuno's theory therefore provides precedent for the rules we give here; a more detailed comparison of the theories will be given in the presentation with time permitting.

[3] Actually, for example (7) the more salient referent *Terry's parents* is ruled out by number agreement.

[4] Ordinarily, the reflexive form could exist if it could be coreferential with the *full* subject NP assuming agreement constraints were met, but this is superfluous to the present discussion.

[5] Actually, intrasentential rules can apply to the subject and entities later in the sentence, as in:

John talked to Bob today.

He usually can't stand to be in the same room as John.

(8) Terry got beaten up by Sandy at school today.
His mother called his father to complain.

The theory also explains why it is difficult to get different referents for pronouns in conjoined NPs:

(9) Terry's parents have been socializing with Bob's parents lately.
Yesterday his mother and his father went to the movies.

As with examples (6)-(8), hearers have a difficult time assigning *Bob* as the referent of the second pronoun after assigning *Terry* as the referent of the first. This is predicted since neither possessive c-commands the other in the conjoined phrase.

The predictions of the theory are further supported by example (10), taken from an actual movie review [9]:[6]

(10) Robert Downey, Jr. does a wonderful job as Chaplin.
I even began to think of him as Chaplin midway through the film.

What is interesting about this example is the use of *Chaplin* instead of *him* in the second sentence. Note that if *him* was used, one is inclined (at least initially) to think that both instances of *him* refer to *Robert Downey Jr.*, the most accessible entity. This is exactly what the theory predicts: since the two pronouns would not stand in the c-command relation, they both would have *Robert Downey Jr.* as their preferred referent (despite the resulting semantic oddness of this assignment in contrast to the alternative of allowing the second *him* to refer to *Chaplin*, which makes perfect sense). Hence, the theory predicts the use of a name instead of a pronoun in this case.

Like all theories of pronoun reference resolution, there are likely to be counterexamples to this theory in actual discourses. However, the unexplained properties of the data presented here suggest that something akin to the proposed constraints is in fact operating. Further work is necessary to adequately evaluate the robustness of the proposed rules.

5 Conclusion

In this abstract I strengthen the ability of Centering Theory to predict correct referents of pronouns without the use of Rule 2. Strong biases for pronominal references in several heretofore unnoted examples are predicted, as are the lack of these biases in other, similar examples. The idea that all

Here, the referent of *He* must be *Bob* due to intrasentential coreference constraints. However, readers generally garden-path on this example, indicating that *John* was the initial entity assigned to *He*.

[6] In keeping with the data-oriented theme of this workshop, a set of over 100 movie reviews was analyzed for data relating to this analysis. Such data is scarce, in that useful sentences have to contain two (potential) pronouns of like agreement features, have the pronouns not in a c-command relationship, and have two distinct potential referents on the preceding Cf list as in example (10). In general, however, I have found movie reviews to be productive sources for considering approaches to pronoun resolution since they often have several actors/characters being discussed at once. This corpus is on-line and available to those requesting it.

pronouns have a preferred referent has strength in its simplicity; furthermore, the rules for treating all pronouns as ambiguous between reflexive and non-reflexive forms has precedent in the Kuno's theory of intrasentential reference [7]. The data presented cast further doubt on proposals that coreference is determined solely on the basis of coherence determination [5, 11].

Acknowledgments

This work was supported in part by National Science Foundation Grant IRI-9009018, National Science Foundation Grant IRI-9157996, and a matching grant for the latter from the Xerox Corporation. I'd like to thank Barbara Grosz and the members of the Harvard AI Research Group for helpful discussions.

References

[1] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. "A Centering Approach to Pronouns", in *Proceedings of the 25th Meeting of the Association for Computational Linguistics*, pp. 155-162, 1987.

[2] Noam Chomsky. *Lectures on Government and Binding*, Foris, Dordrecht. 1981.

[3] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Providing a Unified Account of Definite Noun Phrases in Discourse. In *Proceedings of the 21st Meeting of the Association for Computational Linguistics*, pp. 44–50, Cambridge, MA 1983.

[4] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Towards a Computational Theory of Discourse Interpretation. Unpublished Manuscript. 1990.

[5] Jerry Hobbs. Coherence and Coreference. *Cognitive Science*, vol 3, pp. 67–90, 1979.

[6] Aravind K. Joshi and Scott Weinstein. Control of Inference: Role of Some Aspects of Discourse Structure – Centering", in *Proceedings of IJCAI-81*, pp. 385–387, 1981.

[7] Susumu Kuno. *Functional Syntax*, Chicago Press, 1987.

[8] Howard Lasnik. "Remarks on Coreference", *Linguistic Analysis*, vol. 2, no. 1, 1976.

[9] Afroditi Michailidi. Film review of *Chaplin*. 1992.

[10] Tanya Reinhart. *Anaphora and Semantic Interpretation.* University of Chicago Press, Chicago, 1983.

[11] Robert Wilensky. *Understanding Goal-Based Stories.* Ph.D. Thesis, Yale Univerity Research Report #140, 1978.