Centering: A Parametric Theory and Its Instantiations

Massimo Poesio*
University of Essex

Rosemary Stevenson†
University of Durham

Barbara Di Eugenio‡
University of Illinois at Chicago

Janet Hitzeman§
MITRE Corporation

Centering theory is the best-known framework for theorizing about local coherence and salience; however, its claims are articulated in terms of notions which are only partially specified, such as "utterance," "realization," or "ranking." A great deal of research has attempted to arrive at more detailed specifications of these parameters of the theory; as a result, the claims of centering can be instantiated in many different ways. We investigated in a systematic fashion the effect on the theory's claims of these different ways of setting the parameters. Doing this required, first of all, clarifying what the theory's claims are (one of our conclusions being that what has become known as "Constraint 1" is actually a central claim of the theory). Secondly, we had to clearly identify these parametric aspects: For example, we argue that the notion of "pronoun" used in Rule 1 should be considered a parameter. Thirdly, we had to find appropriate methods for evaluating these claims. We found that while the theory's main claim about salience and pronominalization, Rule 1—a preference for pronominalizing the backward-looking center (CB)—is verified with most instantiations, Constraint 1–a claim about (entity) coherence and CB uniqueness—is much more instantiation-dependent: It is not verified if the parameters are instantiated according to very mainstream views ("vanilla instantiation"), it holds only if indirect realization is allowed, and is violated by between 20% and 25% of utterances in our corpus even with the most favorable instantiations. We also found a trade-off between Rule 1, on the one hand, and Constraint 1 and Rule 2, on the other: Setting the parameters to minimize the violations of local coherence leads to increased violations of salience, and vice versa. Our results suggest that "entity" coherence—continuous reference to the same entities—must be supplemented at least by an account of relational coherence.

1. Motivations

Centering theory (Joshi and Weinstein 1981; Grosz, Joshi, and Weinstein 1983, 1995; Walker, Joshi, and Prince 1998b) is the component of Grosz and Sidner's overall theory

---

* Department of Computer Science, Wivenhoe Park, Colchester CO4 35Q, U.K. E-mail: poesio@essex.ac.uk.
† Department of Psychology, University of Durham, U.K.
‡ Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, USA. E-mail: bdieugen@cs.uic.edu
§ MITRE Corporation, 202 Burlington Road, Bedford, MA 01730-1428, USA. E-mail:hitze@mitre.org.

of attention and coherence in discourse (Grosz 1977; Sidner 1979; Grosz and Sidner 1986) concerned with *local* coherence and salience, that is, coherence and salience within a discourse segment. A fundamental characteristic of centering is that it is better viewed as a *linguistic* theory than a computational one. By this we mean that its primary aim is to make cross-linguistically valid claims about which discourses are easier to process, abstracting away from specific algorithms for anaphora resolution or anaphora generation (although many such algorithms are based on the theory). The result is a very different theory from those one usually finds in computational linguistics. In central papers such as Grosz, Joshi, and Weinstein (1995), no algorithms are provided to compute notions such as "utterance," "previous utterance," "ranking," and "realization" that play a crucial role in the theory. The researchers working on centering argue that while these concepts play a central role in any theory of discourse coherence and salience, their precise characterization is best left for subsequent research, indeed, that some of these concepts (e.g., ranking) might be defined in a different way for each language (Walker, Iida, and Cote 1994). In other words, these notions should be viewed as **parameters** of centering. This feature of the theory has inspired a great deal of research attempting to specify centering's parameters for different languages (Kameyama 1985; Walker, Iida, and Cote 1994; Di Eugenio 1998; Turan 1998; Strube and Hahn 1999). Competing versions of the central definitions and claims of the theory have also been proposed: For example, different definitions of **backward-looking center** (CB) can be found in Grosz, Joshi, and Weinstein (1983, 1995) and Gordon, Grosz, and Gillion (1993). As a result, a researcher wishing to test the predictions of centering, or to use it for practical applications, is confronted with a large number of possible **instantiations** of the theory.

The main goal of the work reported in this article was to explore the space of parameter configurations, measuring the impact of different ways of setting the parameters of centering on the theory's claims. This required specifying in an explicit way what centering's main claims are; clearly identifying the parameters, not all of which have previously been discussed in the literature; and developing appropriate methods (and statistical tests) to carry out this evaluation. The comparison between instantiations was carried out by annotating a corpus of English texts from different genres with the information needed to test a variety of centering instantiations and using this corpus to assess the extent to which the theory's claims are verified once the parameters are set in a certain way. The proponents of centering have clearly stated that the aim of the theory is to identify preferences that make discourses easier to process; clearly, the best way to test such preferences is through behavioral experiments, and many aspects of the theory have in fact been tested this way (Hudson, Tanenhaus, and Dell 1986; Gordon, Grosz, and Gillion 1993; Brennan 1995). But given the enormous number of possible ways of setting the theory's parameters, a systematic comparison can be made only by computational means. A corpus-based evaluation has other advantages, as well, among which is that it is perhaps the best way to identify the aspects of the theory that need to be further specified, and the factors such as temporal coherence or stylistic variation that may interact with the preferences expressed by centering. (Also, knowing the extent to which real texts conform to centering preferences is an important goal in its own right.)

In previous corpus-based studies of centering (Walker 1989; Passonneau 1993; Byron and Stent 1998; Di Eugenio 1998; Kameyama 1998; Strube and Hahn 1999; Tetreault 2001), only a few instantiations of centering were compared. The present study is more systematic in that it considers a greater number of parameters, as well as more parameter instantiations, including "crossing" instantiations in which the parameters are set according to proposals due to different researchers. Only re-

liable annotation techniques were used; we produced an annotation manual that can be used to extend our analysis to other data, as well as a companion Web site (http://cswww.essex.ac.uk/staff/poesio/cbc/) to allow readers to try out instantiations not discussed in this article. (The Web site also contains the annotation manual and a technical report with a full discussion of all results.) Last but not least, our evaluation is arguably more neutral than in most previous studies in that, first of all, we are not proposing a new instantiation of the theory; and secondly, all parameter instantiations were tested on the same data.

The article is organized as follows. We first review the basic concepts of the theory, discussing the three claims on which we focus (Constraint 1, Rule 1, and Rule 2) and the parameters used in their formulation. We then discuss how our corpus was annotated and how the annotation was used to compute violations of the three main claims. In Section 4 we present our main results, which are discussed in Section 5.

2. Centering Theory and Its Parameters

It is not possible to discuss in this article the entire centering literature; we merely summarize in this section some of this work in enough detail to allow the reader to follow the discussion in the rest of the article. For more details, we refer the reader to classic references such as Grosz, Johsi, and Weinstein (1995) and Walker, Joshi, and Prince (1998b) or the discussion of centering in Poesio and Stevenson (forthcoming).

2.1 Motivations and Main Intuitions

Centering is simultaneously a theory of discourse **coherence** and of discourse **salience**. As a theory of coherence, it attempts to characterize **entity-coherent** discourses: discourses that are considered coherent because of the way discourse entities are introduced and discussed.[1] At the same time, centering is also intended to be a theory of salience: that is, it attempts to predict which entities will be most salient at any given time.

The main claim about local coherence made in centering is that discourse segments in which successive **utterances** keep mentioning the same discourse entities are "more coherent" than discourse segments in which different entities are mentioned. This hypothesis was formulated by Chafe (1976) and is backed by empirical evidence such as Kintsch and van Dijk (1978) and Givon (1983). In centering this hypothesis is further strenghtened by proposing that every utterance has a unique "main link" with the previous utterance: the CB. Having a unique CB, it is claimed, considerably simplifies the complexity of the inferences required to integrate an utterance into the discourse (Joshi and Kuhn 1979; Joshi and Weinstein 1981).

Centering's first contention as far as local salience is concerned is that the discourse entities **realized** by an utterance (more on realization below) are *ranked*: that is, that in each utterance some discourse entities are more salient than others. This claim, as well, is a basic tenet of much work on discourse (Sidner 1979; Prince 1981; Givon 1983; Gundel, Hedberg, and Zacharski 1993) and is supported by much psychological evidence (Hudson, Tanenhaus, and Dell 1986; Gernsbacher and Hargreaves 1988; Gordon, Grosz, and Gillion 1993; Stevenson, Crawley, and Kleinman 1994).

---

1 Entity-based theories of coherence are so-called by contrast with **relation-centered** theories of coherence, such as those developed in Hobbs (1979) and Mann and Thompson (1988) and used in Fox (1987) and Lascarides and Asher (1993). The earliest detailed entity-based theory of coherence we are aware of is by Kintsch and van Dijk (1978), who also explicitly mention the need to supplement such theories with a theory of relational coherence (more on this in Section 5)

These claims about coherence and salience are linked by two further hypotheses: that the identity of the CB is crucially determined by the entities' ranking and that the CB is most likely to be realized as a pronoun. This assumption that a "main entity" or "topic" or "focus" is the preferred interpretation of pronouns is commonly found in theories in the psychological (e.g., Sanford and Garrod 1981), computational (Sidner 1979) and linguistic (Gundel, Hedberg, and Zacharski 1993) literature and is motivated by evidence such as the contrast between examples (1) and (2):

(1)   a.   Something must be wrong with John.
        b.   He has been acting quite odd. (He = John)
        c.   He called up Mike yesterday.
        d.   John wanted to meet him quite urgently.

(2)   a.   Something must be wrong with John.
        b.   He has been acting quite odd. (He = John)
        c.   He called up Mike yesterday.
        d.   He wanted to meet him quite urgently.

Discourses (1) and (2) differ only only in their (d) sentences, but according to Grosz et al., (1d) is not as felicitous as (2d). The reason, they argue, is that after the (c) utterances, the discourse entity *John* is more highly ranked than *Mike*, so it will be the CB of the next utterance provided that it is realized in it; and given the preference for pronominalizing the CB, *John* should be pronominalized if anything else is.

This link between pronominalization and the identity of the CB has been used by Grosz et al. to support the claim discussed above that utterances have a unique CB (contra, e.g., Sidner [1979] whose theory assumed two foci). Grosz et al. note the contrast between continuations (3c)–(3f) of the discourse initiated by utterances (3a)–(3b):

(3)   a.   Susan gave Betsy a pet hamster.
        b.   She reminded her that such hamsters were quite shy.
        c.   She asked Betsy whether she liked the gift.
        d.   Betsy told her that she really liked the gift.
        e.   Susan asked her whether she liked the gift.
        f.   She told Susan that she really liked the gift.

Grosz et al. argue that continuations (3c)–(3f) are less and less acceptable, whereas if *Susan* and *Betsy* were equally ranked after (3b), all variants should be equally acceptable.

2.2 Terminology and Definitions

2.2.1 Local Focus, Forward-Looking Centers, and Utterances. A fundamental assumption underlying centering is that processing a discourse involves continuous updates to the local attentional state, or **local focus**. The local focus includes a set of **forward-looking centers** (CFs), which correspond to Sidner's (1979) "potential discourse foci" and can be viewed as mentions of discourse entities (Karttunen 1976; Webber 1978; Heim 1982; Kamp and Reyle 1993). The local focus also contains information about the relative prominence or **rank** of these CFs. The local focus gets updated after every **utterance**: In this update, the current CFs are replaced by new ones, and the CB changes,

as well (see below).[2] The set of CFs introduced in the local focus by utterance U_i in a discourse segment (DS) is indicated by $CF(U_i,DS)$, generally abbreviated to $CF(U_i)$. Brennan, Friedman, and Pollard (1987) formalized the relationship between utterances and CFs by means of one of their so-called constraints:[3]

Constraint 2: Every element of CF(U,DS) must be realized in U.

2.2.2 Ranking, Preferred Centers, and Backward-Looking Centers. We already mentioned two important claims of the theory: that forward-looking centers are ranked, and that because of this ranking, some CFs acquire particular prominence. The ranking function is only required to be partial, but the most highly ranked CF realized by an utterance (when one exists) is called the **Preferred Center** (CP). Ranking is also used to characterize one of the CFs as the CB. The CB is the closest concept in centering to the traditional notion of "topic" (Sgall 1967; Chafe 1976; Sanford and Garrod 1981; Givon 1983; Vallduvi 1990; Gundel, Hedberg, and Zacharski 1993) and plays a central role in the theory's claims about both coherence and salience. Although in Grosz, Joshi, and Weinstein (1983), the CB was characterized only in intuitive terms, most subsequent work has been based on the definition below (Grosz, Joshi, and Weinstein 1995), referred to as "Constraint 3" by Brennan, Friedman, and Pollard (1987):

Constraint 3 $CB(U_i)$, the backward-looking center of utterance U_i, is the highest-ranked element of $CF(U_{i-1})$ that is realized in U_i.

Notice that according to this definition, the computation of the CB depends exclusively on ranking and "previous utterance," making these parameters crucially important for the framework.[4]

2.2.3 Transitions. The hypothesis that discourses are easier to process when successive utterances are perceived as being "about" a unique discourse entity is formalized in centering in terms of a classification of utterances according to the type of **transition** (update) they induce in the local focus. Many such classifications of transitions have been proposed; Grosz, Joshi, and Weinstein (1995) distinguish among three types of transitions, depending on whether the backward-looking center of U_{i-1} is maintained or not in U_i and on whether $CB(U_i)$ is also the most highly ranked entity (CP) of U_i:

Center Continuation (CON): $CB(U_i) = CB(U_{i-1})$, and $CB(U_i)$ is the most highly ranked CF (CP) of U_i (i.e., $CP(U_i) = CB(U_i)$)

Center Retaining (RET): $CB(U_i) = CB(U_{i-1})$, but $CP(U_i) \neq CB(U_i)$

Center Shifting (SHIFT): $CB(U_{i-1}) \neq CB(U_i)$

We will consider a few alternative classification schemes below, after discussing how these classifications are used to formulate one of the core claims of centering, Rule 2.

---

2 The hypothesis that discourse processing involves continuous updates to the discourse model also lies at the heart of so-called "dynamic" theories of discourse semantics (Heim 1982; Kamp and Reyle 1993).
3 The order of presentation of constraints and rules followed here differs from that more familiar in the centering literature. This is because we want to distinguish between definitions and claims, and the three constraints proposed by Brennan et al. do not all have the same status: While Constraint 2 can be seen as a "filter" ruling out certain values of $CF(U_i)$, Constraint 3 is a definition, and Constraint 1 an empirical claim.
4 Other ways of defining the CB have been proposed. We refer the interested reader to the longer report and to the companion Web site.

2.3 Main Claims
In the words of Grosz et al., the most fundamental claim of centering is that "to the extent that discourse adheres to centering constraints, its coherence will increase and the inference load placed upon the hearer will decrease" (Grosz, Joshi, and Weinstein 1995, page 210). Grosz et al. list seven such "costraints," three of which can be directly evaluated. Even though we are not following here the distinction between "constraints" and "rules" introduced in Brennan, Friedman, and Pollard (1987), we will use for these three claims the names Brennan et al. gave them, by which they are now best known:

Constraint 1 (Strong): All utterances of a segment except for the first have exactly one CB.

Rule 1 (GJW95): If any CF is pronominalized, the CB is.

Rule 2 (GJW95): (Sequences of) continuations are preferred over (sequences of) retains, which are preferred over (sequences of) shifts.

2.3.1 Constraint 1, Topic Uniqueness, and Entity Coherence. If we view the CB as a formalization of the idea of "topic" (Vallduvi 1990; Gundel 1998; Hurewitz 1988; Miltsakaki 1999; Beaver 2004), Constraint 1 expresses, first and foremost, the original claim from Joshi and Kuhn (1979) and Joshi and Weinstein (1981) that discourses with exactly one (or no more than one) "topic" at each point are easier to process. This view contrasts both with Sidner's (1979) hypothesis that utterances may have two "topics" and with theories such as Givon (1983), Alshawi (1987), Lappin and Leass (1994) and Arnold (1998), which view "topichood" as a matter of degree and therefore allow for an arbitrary number of topics.

In the strong form just presented, Constraint 1 is also a claim about local coherence. It expresses a preference for discourses to be entity coherent: to continue talking about the same entities. Each utterance in a segment should realize at least one of the discourse entities realized in the previous utterance. A weaker form of Constraint 1 has also been suggested (e.g., Walker, Joshi, and Prince 1998a, footnote 2, page 3); the preference for a unique CB is preserved, but not the preference for entity coherence.

Constraint 1 (Weak): All utterances of a segment except for the first have *at most one* CB.

2.3.2 Rule 1 and Pronominalization. Rule 1 is the main claim of centering about pronominalization. In the version presented above, it states a preference for pronominalizing the CB, if anything is pronominalized at all. We also examined two alternative formulations. The original form of the claim in Grosz, Joshi, and Weinstein (1983) was as follows:

Rule 1 (GJW83): If the CB of the current utterance is the same as the CB of the previous utterance, a pronoun should be used.

Gordon, Grosz, and Gillion (1993) proposed a much stronger form of the claim. They found that entities realized in certain positions in the sentence were read more slowly unless pronominalized (**repeated-name penalty** [RNP]).[5] This evidence led them to

---

5 Gordon et al. observed increased reading times when proper names were used instead of pronouns to realize an entity in subject position referring to an entity realized in first-mentioned or subject position.

propose a more restrictive definition of CB (briefly, that the CB is the entity subject to the RNP—for discussion, see the longer version of this article available on the Web site) and a stronger form of Rule 1, requiring the CB (defined in this more restrictive way) to be always pronominalized:

Rule 1 (Gordon et al.): The CB should be pronominalized.

Although we will refer to this version as "Gordon et al.'s" for brevity, readers should keep in mind that because the definition of CB proposed by Gordon et al. is more restrictive, their version of Rule 1 is properly evaluated only using that definition. (The results with this instantiation are discussed in the longer version of the article availabel on the Web site.)

2.3.3 Rule 2 and the Classification of Transitions. Rule 2 is a claim about coherence, as well: It states a preference for preserving the CB over changing it and for realizing it as the most salient entity over changing its relative ranking. This aspect of the theory has received a lot of attention; several variants of this constraint have been proposed, as well as many ways of classifying transitions. We studied many alternative proposals.

The version of Rule 2 presented in Grosz, Joshi, and Weinstein (1995) expresses preferences among *sequences* of transitions (e.g., CON-CON over SHIFT-SHIFT) rather than preferences for particular transitions. This form of the constraint is in part motivated by empirical results. Di Eugenio (1998), for example, found that the relative distribution of null and explicit pronouns in Italian depends on the previous transition as well: in Center Continuations that follow a CON or a SHIFT, it is much more likely that a null pronoun will be used, whereas in Center Continuations that follow a RET transition, both null and explicit pronouns are equally likely. Turan (1998) obtained similar results for null and explicit pronouns in Turkish.

Other researchers argue instead that the inferential load is evaluated utterance by utterance (Brennan, Friedman, and Pollard 1987; Walker, Iida, and Cote 1994; Walker, Joshi, and Prince 1998a). The version of Rule 2 proposed by Brennan et al. is as follows:

Rule 2 (Single Transitions): Transition states are ordered. The CON transition is preferred to the RET transition, which is preferred to the Smooth Shift transition, which is preferred to the Rough Shift transition.

This formulation of Rule 2 depends on a further distinction between two types of SHIFT: Smooth Shift (SSH), when $CB(U_n) = CP(U_n)$, and Rough-Shift (RSH), when $CB(U_n) \neq CP(U_n)$. Transitions can then be classified along two dimensions, as in the following table:

	$CB(U_n) = CB(U_{n-1})$ or $CB(U_{n-1}) = $ NIL	$CB(U_n) \neq CB(U_{n-1})$
$CB(U_n) = CP(U_n)$	Continue	Smooth Shift
$CB(U_n) \neq CP(U_n)$	Retain	Rough Shift

Further refinements of these classification schemes have been proposed. Kameyama (1986) proposed a fourth transition type, Center Establishment (EST), for utterances

---

E.g., in *Bruno was the bully of the neighborhood. Bruno / He often taunted Tommy*, the second sentence would be read more slowly when *Bruno* was used than when *He* was used.

that establish a CB after an utterance without one, such as the first utterance of a segment. Walker, Iida, and Cote (1994) argued that these utterances should be classified as Center Continuations, the idea being that even the first utterance of a segment does have a CB, but this CB is initially underspecified and is determined only when the second utterance is processed. Notice that according to the strong version of Constraint 1, the first utterance of a discourse segment is the only utterance allowed not to have a CB in a coherent discourse; hence, none of these classification schemes for transitions includes classes either for the inverse of Center Establishment, that we might call Zeroing (ZERO) transition (a CB less utterance following one which does have a CB) or for CB-less utterances following other CB-less ones (the NULL transition).

Strube and Hahn (1999), like Grosz, Joshi, and Weinstein (1995), claim that inferential load is evaluated across sequences (pairs, in fact) of transitions but argue for a different way of evaluating the inferential load of utterances. In their view, classifications of transitions such as those above do not reflect what should be one of the crucial claims of the theory: that the CP of one utterance predicts the CB of the next. In order to formalize their view, they propose a different classification scheme, based on the distinction between **cheap** and **expensive** transitions (Strube and Hahn 1999, page 332):

- A transition pair is **cheap** if the CB of the current utterance is correctly predicted by the CP of the previous utterance, that is, if $CB(U_n) = CP(U_{n-1})$.

- A transition pair is **expensive** if $CB(U_n) \neq CP(U_{n-1})$.

Strube and Hahn then propose a new version of Rule 2 based on this distinction:

Rule 2 (Strube and Hahn): Cheap transition pairs are preferred to expensive ones.[6]

2.4 The Parameters of Centering
Although Grosz et al. discussed possible definitions for the concepts used in the claims above–"utterance," "previous utterance," "ranking," and "realization"—they didn't settle on a specific definition, even for English. Similarly undefined is the notion of "pronominalization" governed by Rule 1. But without further specification of these concepts it is impossible to evaluate the claims above, just as it is not possible to evaluate the predictions of, say, "government and binding theory" without providing an explicit definition of "command" or "argument". As a result, a considerable amount of research has been concerned with establishing the best specification for what are, essentially, parameters of the theory. We briefly review some of these proposals in this section.[7]

---

6 Kibble (2001) proposed a version of Rule 2 that further develops the "decompositional" view of Rule 2 introduced by Brennan et al., while simultaneously incorporating Strube and Hahn's intuition that "cheap" transitions should be preferred. Kibble formulates his version of Rule 2 as a series of preferences: for transitions that preserve the CB—that is, those such that $CB(U_n) = CB(U_{n-1})$ (he calls these transitions cohesive), that identify $CB(U_n)$ with $CP(U_n)$, and/or that are cheap. Code to test an earlier version of Kibble's form of Rule 2 (Kibble 2000) has been incorporated in the scripts discussed later in the article, and the results can be seen in the longer technical report accompanying this article, or from the companion Web site. We will, however, omit a discussion of this version here, in part for reasons of space, in part because the final version of the rule in Kibble (2001) differs substantially from the original version that we evaluated.
7 For more details, and for a discussion of the motivations behind these proposals, see the extended version of this article and Poesio and Stevenson (forthcoming).

2.4.1 Utterance and Previous Utterance. In the early centering papers, utterances were implicitly identified with sentences. Kameyama (1998), however, argued that such identification makes the number of potential antecedents of anaphoric expressions much greater than if they were resolved clause by clause. Furthermore, she noted that this identification leads to problems with multiclausal sentences: for example, grammatical function ranking becomes difficult to compute, as a sentence may have more than one subject. Kameyama proposed that the local focus be updated after every tensed clause, not after every sentence, and classified tensed clauses into (1) utterances that constitute a "permanent" update of the local focus, such as coordinated clauses and adjuncts, and (2) **embedded** utterances that result in temporary updates that are then "popped", much as the information introduced into discourse by subordinated discourse segments is popped according to Grosz and Sidner (1986). According to Kameyama, only a few types of clauses, such as the complements of certain verbs, are embedded. For example, Kameyama proposes to break up (4) into utterances as follows, and to treat each of these utterances, including subordinate clauses such as (U2) or (U5), as an update:

(4)   (u1) **Her** entrance in Scene 2 Act 1 brought some disconcerting applause
        (u2) even before **she** had sung a note. (u3) Thereafter the audience
        waxed applause happy (u4) but discriminating operagoers reserved
        judgment (u5) as **her** singing showed signs of strain.

Experiments by Pearson, Stevenson, and Poesio (2000) confirmed that CFs introduced in main clauses are significantly more likely to be subsequently mentioned than CFs introduced in complement clauses. However, a semicontrolled study by Suri and McCoy (1994) suggested that other types of clauses—specifically, adjunct clauses headed by *after* and *before*–are also "embedded," not "permanent updates" as suggested by Kameyama; these results were subsequently confirmed by Cooreman and Sanford (1996). The status of other types of clauses is less clear. Kameyama (1998) also proposed a tentative analysis of relative clauses, according to which they are temporarily treated as utterances and update the local focus but are then merged with the embedding clause; she didn't, however provide empirical support for this hypothesis. Other types of subordinate clauses and parentheticals are not discussed in this literature.

Strube (1998) and Miltsakaki (1999) question Kameyama's identification of utterances with (tensed) clauses. Miltsakaki (1999) argues, on the basis of data from English and Greek, that the local focus is updated only after every sentence and that only the CFs in the main clause are considered when establishing the CB.

2.4.2 Realization. Grosz, Joshi, and Weinstein (1995) simply say that what it means for utterance U to realize center C depends on the particular semantic theory one adopts. They consider two ways in which a discourse entity may be "realized" in an utterance as required by Constraint 2. **Direct** realization is when a noun phrase in the utterance refers to that discourse entity. **Indirect** realization is when one of the noun phrases in the utterance is an **associative reference** to that discourse entity in the sense of Hawkins (1978),[8] that is, an anaphoric expression that refers to an object which wasn't mentioned before but is somehow related to an object that already has. For example, in the following discourse:

(5)   (u1) John walked toward *the house*. (u2) <u>The door</u> was open.

---

8 Associative references are one type of **bridging reference** (Clark 1997).

John, *the house* and *the door* are directly realized in the respective utterances; in addition, *the house* can be thought as being indirectly realized in (u2) by virtue of being referred to by the associative reference *the door* (see, e.g., the discussion in Grosz, Joshi, and Weinstein [1995] and Walker, Joshi, and Prince [1998b]). Clearly, the computation of the CB is affected by which entities are considered to be "realized" in an utterance: In (5), for example, (u2) has a CB (the house) only if *the house* is considered to be realized in (u2) by virtue of its being associated with *the door*. To our knowledge, the effect of these alternative notions of realization on the predictions of the theory has not been previously studied, even though theories of focusing such as Sidner's (1979) do allow the (discourse) focus to be realized in an utterance in these cases, and the issue is often mentioned in discussions of centering.

A related issue is whether empty realizations, or **traces**, should count as realizations of an entity. Many theories of grammar hypothesize that morphologically null elements occur in the syntactic structure underlying a variety of constructions, including control constructions as in (6a), reduced relatives as in (6b), and even coordinated verb phrases (VP) as in (6c):

(6) a. John wanted ($\emptyset$ to buy a house).
     b. John bought a house ($\emptyset$ abandoned by its previous occupiers).
     c. John bought a house and ($\emptyset$ promptly demolished it).

If, for example, the coordinated VP in (6c) is considered a separate utterance, whether or not it contains a realization of *John* is going to determine whether or not it has a CB. To our knowledge, morphologically null elements have been considered in the centering literature only for languages other than English.

An issue that has been raised in the centering literature (e.g., Walker 1993; Di Eugenio 1998; Byron and Stent 1998) is whether the CF list contains only entities realized as third-person noun phrases (NPs), or also the entities realized as first- and second-person NPs. Walker (1993) suggests that deictic entities are beyond the purview of centering; however, in example (7), neither utterance (u2) nor utterance (u3) would have a CB if second-person pronoun *you* is not counted as introducing an entity in the CF list.[9]

(7) (u1) You should not use PRODUCT-Z
     (u2) if you are pregnant of breast-feeding.
     (u3) Whilst you are receiving PRODUCT-Z ...

2.4.3 Ranking. Perhaps the most discussed parameter of centering, at least in the versions of the theory that accept the definition of CB specified by Constraint 3, is the ranking function. Most researchers working on centering, including Grosz et al., assume that several factors play a role in determining the relative ranking of forward-looking centers; in fact, Walker, Iida, and Cote (1994) and Walker, Joshi, and Prince (1998a) claim that the factors affecting ranking may not be the same in all languages.

---

9 According to Walker, Joshi, and Prince (1998a), in the original version of Grosz, Joshi, and Weinstein (1995), which appeared in 1986, Grosz et al. provided a more explicit definition of realization:

  An utterance U realizes a center c if c is an element of the situation described by U, or c is the semantic interpretation of some subpart of U.

  With this definition, all of the cases considered above—the anchors of associative references, traces, and the entities realized as first- and second-person pronouns—would be considered as realized by an utterance.

Nevertheless, most versions of the theory developed since Kameyama (1985, 1986) and Grosz, Joshi, and Weinstein (1986) have assumed that **grammatical function** plays the main role in determining the order among forward-looking centers, at least for English. Specifically, Grosz, Joshi, and Weinstein (1995) claim that subjects are ranked more highly than objects, which are ranked more highly than other grammatical positions: SUBJ ≺ OBJ ≺ OTHERS (see also Kameyama 1986; Hudson, Tanenhaus, and Dell 1986). Slightly different ranking functions based on grammatical function were proposed by Brennan, Friedman, and Pollard (1987) (who made a further distinction between objects and indirect objects), by Walker, Iida, Cote (1994) for Japanese, and by Turan (1998) for Turkish. There is quite a lot of psychological support for at least the part of this claim stating that entities realized as subjects are more salient than entities realized in other grammatical functions (Hudson, Tanenhaus, and Dell 1986; Gordon, Grosz, and Gillion 1993; Brennan 1995; Hudson-D'Zmura and Tanenhaus 1998).

Other factors affecting ranking have been considered as well. Rambow (1993) proposed that a number of facts about scrambling in German could be explained if ranking in German were to be determined by surface **order of realization**. The idea that order of mention affects salience is well supported by psychological evidence; for example, the results of probe experiments by Gernsbacher and Hargreaves (1998) suggest that the first-mentioned discourse entity in a sentence is the most salient. The interaction of order of mention with grammatical function has also been studied. As mentioned above, Gordon, Grosz and Gillion (1993) observed a repeated name penalty for CFs in subject position coreferring with an entity previously introduced. This effect was observed both when the antecedent was in subject position and when it was the first-mentioned entity in a nonsubject position (as in *In Lisa's opinion, he shouldn't have done that*), suggesting that first-mentioned CFs are as highly ranked as subjects.

Strube and Hahn (1999) argue that in German, the rank of discourse entities is determined by the position they hold in Prince's (1981, 1992) givenness hierarchy. Specifically, Strube and Hahn argue that hearer-old entities rank higher than mediated entities, which and in turn rank higher than hearer-new entities: HEARER-OLD ≺ MEDIATED ≺ HEARER-NEW.[10] Order of mention also plays a role in their ranking: Within each category, the entities realized earlier in the sentence are ranked more highly.

Finally, Sidner's original (1979) claim that ranking depended on thematic roles, abandoned in the early versions of centering, was revisited by Cote (1998). This view is supported by psychological work on "implicit causality" verbs (Caramazza et al. 1997) as well as work by Stevenson, Crawley, and Kleinman (1994), and Pearson, Stevenson, and Poesio (2001). In particular, there is evidence that with certain verbs, the normal preference for subjects to rank higher than their objects is reversed, although these preferences are modified by other factors such as order of mention, type of connective, and animacy (Stevenson, Crawley, and Kleinman 1994; Stevenson et al. 2000; Pearson, Stevenson, and Poesio 2001).

2.4.4 R1-Pronouns. Rule 1 states that if any CF is pronominalized, the CB is, but the theory does not explicitly specify which types of "pronouns" are covered by this rule. It seems clear that realization as a third-person singular pronoun does count; that is, if the choice is between using a third-person singular pronoun to realize a

---

10 Strube and Hahn's hearer-old entities include Prince's evoked (= discourse old) and unused entities, which are entities such as *Margaret Thatcher* that are supposed to be part of shared knowledge. Mediated entities are the entities falling in Prince's categories inferrable, containing inferrable, and anchored brand-new.

CB or another CF, the CB should be chosen. We also saw that in languages such as Italian, Japanese, and Turkish, the preferred realizations of CBs are morphologically null elements (Kameyama 1986; Walker, Iida, Cote 1994; Turan 1998; Di Eugenio 1998). But should an utterance in English count as verifying the rule if a CF is realized as a third-person pronoun, and the CB as a trace? Or if the CB is realized with a full NP, but a second CF is realized with a demonstrative pronoun? And what about first- and second-person pronouns? The precise characterization of the (sub)class of pronouns subject to Rule 1, which we will call **R1-pronoun**s, is clearly an essential aspect of the theory, yet to the best of our knowledge, no proposals in this regard can be found in the centering literature.

2.5 Empirical Support for, and Applications of, Centering

Centering has served as the theoretical foundation for a lot of work in linguistics, natural language processing (NLP), and psychology. This includes annotation studies testing the claims of the theory for languages including English, German, Hindi, Italian, Japanese, and Turkish (e.g., Kameyama 1985; Passonneau 1993; Walker, Iida, and Cote 1994; Di Eugenio 1998; Turan 1998) and several papers in (Walker, Joshi, and Prince (1998b). The claims about pronominalization made in centering have been applied to develop algorithms both for anaphora resolution (Brennan, Friedman, and Pollard 1987; Strube and Hahn 1999; Tetreault 2001) and for sentence planning (Dale 1992; Henschel, Cheng, and Poesio 2000); this work can be viewed as providing an evaluation of claims such as Rule 1. Ideas from centering, and in particular Rule 2, are increasingly found useful in text planning (McKeown 1985; Kibble and Power 2000; Knott et al. 2001; Karamanis 2003).

We have already seen that some predictions of the theory have also been tested with psychological techniques. In many of these experiments, differences in processing pronominal references in a sentence in a sentence to entities with different ranks (according to a particular instantiation of the theory) were observed: Hudson, for example, observed that pronominal references to entities introduced in subject position in the previous sentence are interpreted more quickly than nonpronominal references or references to nonsubjects (Hudson, Tanenhaus, and Dell 1986; Hudson-D'Zmura and Tanenhaus 1998). And we already mentioned that Gordon, Grosz, and Gillion (1993) identified a processing time slowdown, the RNP, when NPs in subject position referring to entities introduced in subject or first-mentioned position in the previous sentence are not pronominalized.

However, the discussion in this section should have made it clear just how many parameters the theory has and in how many different ways they can be instantiated. To our knowledge, no previous study has attempted to analyze in a systematic way how varying the instantiation of more than one of these parameters affects the claims of the theory, especially for combinations of parameter settings not considered in the original papers. This analysis is the goal of the work discussed here.

3. A Corpus-Based Comparison of Centering's Instantiations

Given the many ways in which the parameters of centering can be set, the only feasible way to make a systematic comparison between the theory's "instantiations" is by computational means. That is, running computer simulations of the process of local focus update using an annotated corpus and comparing the results obtained under different instantiations. The evaluation principle we use for this comparison is the number of "violations" of the theory's claims resulting when the parameters are set in a certain way (e.g., whether pronominalization choices are in accordance with Rule

1). In this section we discuss how we set about performing such comparison, the data we used, our annotation methods, and how the annotation was used.

3.1 Evaluating the Claims of Centering against a Corpus

A preliminary question we had to address is what are in fact the main claims of the theory. As discussed in Section 2, of the seven claims mentioned in Grosz, Joshi, and Weinstein (1995), Constraint 1, Rule 1, and Rule 2 are the ones that can actually be verified using a corpus; we concentrated on these. Because several variants of these three claims have been proposed, we evaluated a few of these variants as well.[11]

The second important question is how these three claims are meant to be interpreted and what we can expect a corpus to tell us about them. The proponents of centering are quite clear that the theory does not state "hard" facts about language, that is, the kind of facts whose violation leads to ungrammaticality judgments. Constraint 1, Rule 1, and Rule 2 are meant to be preferences which, when followed, lead to texts that are easier to process.[12] The mere presence of a few exceptions to a claim does not, therefore, count as a falsification. For one thing, we should expect centering preferences to interact with other constraints (a point not emphasized enough in the centering literature). And for another, there may be no way of expressing a particular piece of information without violating some such preferences.[13] So at best, we can expect the three claims to be verified in a *statistical* sense: that is, that the number of utterances that verify such claims will be significantly higher than the number of utterances that violate them—and in fact, we may find that for some claims, even statistical significance will not be achieved. As a result, we evaluated each claim using statistical tests; the tests we used are the sign test for Constraint 1 and Rule 1, and the Page test for Rule 2 (Siegel and Castellan 1988).

It is also important to keep in mind that a corpus cannot tell us whether "violations" actually result in processing difficulties: This can be determined only by behavioral studies such as reading-time experiments. So we should make it absolutely clear that minimizing violations cannot and should not be the only deciding factor in theorizing about centering. Nevertheless, the combinatorics of the problem make it impossible to compare the parameter instantiations in any other way. Furthermore, this form of evaluation is also the most systematic way to identify other preferences and constraints that may interact with centering. We return to these issues in Section 5.

3.2 The Data

The data used in this work are texts from the GNOME corpus, which currently includes texts from three domains. The museum subcorpus consists of descriptions of museum objects and brief texts about the artists who produced them.[14] The pharmaceutical subcorpus is a selection of leaflets providing patients with legally mandatory

---

11 In this version of the article, we assume that the CB is defined by Constraint 3. For the results with alternative definitions of CB, see the extended technical report or the companion Web site.

12 Beaver (2004) argues—correctly, in our opinion—that in one of the best-known pronoun resolution algorithms based on centering, that proposed by Brennan, Friedman, and Pollard (1987), Rule 1 is effectively used as a hard constraint, a problem fixed by Beaver's own optimality-theoretic reformulation of the algorithm. It is nevertheless quite clear that in the theory, Rule 1 has the status of a preference.

13 This point is especially important from an NLG perspective: see, e.g., Karamanis (2003). We will return to this issue in Section 5.

14 The museum subcorpus extends the corpus collected to support the ILEX and SOLE projects at the University of Edinburgh. ILEX generates Web pages describing museum objects on the basis of the perceived status of its user's knowledge and of the objects she has previously looked at (Oberlander et al. 1998) The SOLE project ILEX with concept-to-speech abilities, using linguistic information to control intonation (Hitzeman et al. 1998).

information about their medicines.[15] The GNOME corpus also includes tutorial dialogues from the Sherlock corpus collected at the University of Pittsburgh (Di Eugenio, Moore, and Paolucci 1997). Each subcorpus contains about 6,000 NPs. Texts from the first two domains were used for the main experiments reported here. The third subcorpus was used for the segmentation experiments discussed in the extended technical report available on the Web site.

The data used for this study have two characteristics that make them of particular interest. First of all, they cover genres not previously considered in studies on centering and more similar to those that "real" NLP applications have to contend with. At the same time, they are strongly entity-centered (see, e.g., Knott et al [2001] for an analysis of the museum data), so the hypotheses about coherence formulated in centering are likely to play an important part in the way these texts are constructed.

3.3 Annotation
The previous corpus-based investigations of centering theory we are aware of (Walker 1989; Passonneau 1993, 1998; Byron and Stent 1998; Di Eugenio 1998; Hurewitz 1998; Kameyama 1998; Strube and Hahn 1999) were all carried out by a single annotator annotating her or his corpus according to her or his own subjective judgment. One of our goals was to use for this study only information that could be annotated reliably (Passonneau and Litman 1993; Carletta 1996), as we believe this will make our results easier to replicate. The price we paid to achieve replicability is that we couldn't test all proposals about the computation of centering parameters proposed in the literature, especially those about segmentation and about ranking, as discussed below. The annotation followed a detailed manual, available on the companion Web site. Eight paid annotators were involved in the reliability studies and the annotation. In the following we briefly discuss the information that we were able to annotate, what we didn't annotate, and the problems we encountered; for more details, we refer readers to the extended version of the article and the Web site.

A systematic comparison among different ways of setting the parameters would be prohibitively expensive with traditional psychological methods, but it's not easy to do with corpus analysis, either. Obviously, it can't be done by directly annotating "utterances" or "CB" according to one way of fixing the parameters, as has been done in most previous studies of centering theory (Byron and Stent 1998; Di Eugenio 1998; Kameyama 1998; Passonneau 1993; Walker 1989). Instead, we annotated our corpus with the primitive concepts used by different instantiations of the theory, that is, information that has been claimed by one or the other instantiation of centering to play a role in the definitions of its basic notions. This includes, for example, how sentences break up into clauses and subclausal units; grammatical function; and anaphoric relations, including bridging references. An automatic script uses this information to compute utterances, their CF ranking, and their CB, according to a particular way of setting the parameters, and to compute statistics relevant to the three claims according to that instantiation.

3.3.1 Utterances. In order to evaluate the definitions of utterance proposed in the literature (sentences versus finite clauses), as well as the different proposals concerning the "previous utterance" discussed above, we marked all spans of text that might be claimed to update the local focus. This includes sentences (defined as all units of text

---

15 The leaflets in the pharmaceutical subcorpus are a subset of the collection of all patient leaflets in the United Kingdom, which was digitized to support the ICONOCLAST project at the University of Brighton, developing tools to support multilingual generation (Scott, Power, and Evans 1998).

ending with a period, a question mark, or an exclamation point) as well as what we called (**discourse**) **units**. Units include clauses (defined as sequences of text containing a verbal complex, all its obligatory arguments, and all postverbal adjuncts) as well as other sentence subconstituents that might independently update the local focus, such as parentheticals, preposed prepositional phrases (PPs), and (the second element of) coordinated VPs.[16]

Sentences have one attribute, `stype`, specifying whether the sentence is declarative, interrogative, imperative, or exclamative. The attributes of units include

- `utype`: whether the unit is a main clause, a relative clause, an appositive, a parenthetical, etc.
- `verbed`: whether the unit contains a verb or not.
- `finite`: for verbed units, whether the verb is finite or not.

Marking up sentences proved to be quite easy; marking up units, on the other hand, required extensive annotator training. The agreement on identifying the boundaries of units, using the kappa statistic discussed in Carletta (1996), was $\kappa = .9$ (for two annotators and 500 units); the agreement on features (two annotators and at least 200 units) was as follows: utype: $\kappa = .76$; verbed: $\kappa = .9$; finite: $\kappa = .81$. In total, the texts used for the main study contain 505 sentences and more than 1,000 units, including 900 finite clauses

3.3.2 NPs. Our instructions for identifying NP markables derive from those proposed in the MATE scheme for annotating anaphoric relations (Poesio, Bruneseaux, and Romary 1999), in turn derived from DRAMA (Passonneau 1997) and MUC-7 (Chinchor and Sundheim 1995). In the GNOME corpus, NPs are marked by <ne> ('Nominal entity') elements. In total, the texts used for this study contain 3,345 NPs. These include 586 pronouns, among which are 217 third-person personal and possessive pronouns, 23 demonstratives, and 308 second-person pronouns; 1,290 definite NPs, including 554 *the*-NPs, 250 possessive NPs, and 391 proper nouns; 1,119 indefinite NPs, including 745 bare NPs and 269 *a*-NPs; and 350 other NPs, including 117 quantified NPs and 114 coordinated NPs.

We annotated 14 attributes of <ne> elements specifying their syntactic, semantic, and discourse properties (Poesio 2000). Those relevant to the study discussed here include the following:

- The NP type, cat, with values such as a-np, that-np, the-np, and pers-pro.
- The agreement features num, per, and gen, used to identify contexts in which the antecedent of a pronoun could be identified unambiguously.
- The grammatical function gf. Our instructions for this feature are derived from those used in the FRAMENET project (Baker, Fillmore, and Lowe 1998); see also http://www.icsi.berkeley.edu/˜framenet/. The values are subj, obj, predicate (used for postverbal objects in copular sentences, such as *This is (a production watch)*), there-obj (for postverbal objects in *there*-sentences), comp (for indirect objects), adjunct (for the argument of PPs modifying VPs), gen (for NPs in determiner position in possessive NPs), np-compl, np-part, np-mod, adj-mod, and no-gf (for NPs occurring by themselves—eg., in titles).

---

16 Our instructions for marking up such elements benefited from the discussion of clauses in Quirk and Greenbaum (1973) and from Marcu's (1999) proposals for discourse units annotation.

The agreement values for these attributes are as follows: `cat`: .9; `gen`: .89; `gf`: .85; `num`: .84; `per`: .9. We encountered problems even with supposedly "easy" information such as number and gender, but especially so with semantic attributes (see the annotation manual available on the Web site). We were, however, able to mark up the attributes relevant for this study in a reliable fashion. One exception is that we weren't able to reach acceptable agreement on a feature of NPs often claimed to affect ranking, thematic roles (Sidner 1979; Cote 1988; Stevenson, Crawley, and Kleinman 1994); the agreement value in this case was $\kappa = .35$. As a result, we were not able to evaluate ranking functions based on thematic roles.

3.3.3 Anaphoric Information. In order to determine whether a CF of an utterance is realized directly or indirectly in the following utterance, it is necessary to annotate the anaphoric relations that CFs enter into, including both identity relations and, in order to compute indirect realization, associative relations (Hawkins 1978). This type of annotation raises, however, a number of difficult (and sometimes unresolved) semantic issues (Poesio 2004). As part of the MATE and GNOME projects, an extensive analysis of previously existing schemes for so-called "coreference annotation," such as the MUC-7 scheme, was carried out, highlighting a number of problems with such schemes, ranging from issues with the annotation methodology to semantic issues. Although some of these schemes, like DRAMA, allow the marking of associative relations, none of them analyze which among such relations can be reliably annotated (Poesio, Bruneseaux, and Romary 1999; Poesio 2000). The semantic problems with these schemes include the inappropriate use of the term *coreference* to cover semantic relations such as that between an intensional entity like *the temperature* that may take different values at different time points and these values (as in *the price of aluminum siding rose from $3.85 to $4.02*) or between a quantifier and a variable the quantifier binds, in which neither may corefer (as in *none of the meetings resulted in an agreement between its participants* (van Deemter and Kibble 2000; Poesio 2004). In MATE, a general scheme was developed which includes a fine-grained repertoire of semantic relations, such as binding and function value (Poesio, Bruneseaux, and Romary 1999). For the GNOME corpus, we adopted a simplified version of the MATE scheme, as for our purposes it's not essential to mark all semantic relations between entities introduced by a text, but only those that may establish a "link" between two utterances. So for example, it is in general unnecessary in our case to mark a relation between the subject of a copular sentence and its predicate (e.g., between *the price of aluminum siding* and either *$3.85* or *$4.02* in the example above). Also, our texts do not include any case of bound anaphora, so it was not necessary to offer this option to our annotators.

In the GNOME corpus, anaphoric information is marked by means of a special ⟨ante⟩ element; the ⟨ante⟩ element itself specifies the index of the anaphoric expression (a ⟨ne⟩ element) and the type of semantic relation (e.g., identity), whereas one or more embedded ⟨anchor⟩ elements indicate possible antecedents.[17] (See (8).)

```
(8)     <unit finite='finite-yes' id='u227'>
        <ne id='ne546' gf='subj'>  The drawing of
          <ne id='ne547' gf='np-compl'>the corner cupboard </ne></ne>
        <unit finite='no-finite' id='u228'>,   or more probably
          <ne id='ne548' gf='no-gf'> an engraving of
              <ne id='ne549' gf='np-compl'>  it </ne></ne>
```

---

17 The presence of more than one ⟨anchor⟩ element indicates that the anaphoric expression is ambiguous.

```
   </unit>,
   .
   .
   .
   </unit>
   <ante current="ne549" rel="ident"> <anchor ID="ne547"> </ante>
```

Work such as Sidner (1979) and Strube and Hahn (1999) as well as our own prelimi-
nary analysis, suggested that indirect realization can play a crucial role in maintaining
the CB. However, previous attempts at marking anaphoric information, particularly
in the context of the MUC initiative, suggested that while it's fairly easy to achieve
agreement on identity relations, marking up associative relations is quite hard; this
was confirmed by studies such as Poesio and Vieira (1998). For these reasons, and to
reduce the annotators' work, we marked only a few types of relations, and we specified
priorities. Besides identity (IDENT), we marked up only three associative relations: set
membership (ELEMENT), subset (SUBSET), and "generalized possession" (POSS), which
includes part-of relations as well as ownership relations. We marked only relations
between objects realized by noun phrases and not, for example, anaphoric references
to actions, events, or propositions implicitly introduced by clauses or sentences. We
also gave strict instructions to our annotators concerning how much to mark. (See
the annotation manual available on the Web site.) Furthermore, we specified prefer-
ences: For example, in *Francois, the Dauphin*, the embedding NP would be chosen as
an antecedent, rather than the NP in appositive position.

    As expected, we found a reasonable (if not perfect) agreement on identity relations.
In our most recent analysis (two annotators marking the anaphoric relations between
200 NPs) we observed no real disagreements; 79.4% of the relations were marked up
by both annotators; 12.8% by only one of them; and in 7.7% of the cases, one of the
annotators marked up a closer antecedent than the other.[18] With associative references,
limiting the relations did curtail the disagreements among annotators (only 4.8% of the
relations are actually marked differently), but only 22% of bridging references were
marked in the same way by both annotators; 73.17% of relations were marked by only
one or the other annotator. So reaching agreement on this information involved several
discussions between annotators and more than one pass over the corpus (Poesio 2000).

3.3.4 Segmentation. According to Grosz and Sidner (1986), centering is meant to cap-
ture preferences only within discourse segments. A proper evaluation of the claims of
the theory would therefore require a corpus in which discourse segments have been
identified. Unfortunately, discourse segments are difficult to identify reliably (Passon-
neau and Litman 1993), and Grosz and Sidner (1986) do not provide a specification
of discourse intentions explicit enough that it can be used to identify the intentional
structure of texts—which, according to Grosz and Sidner, determines their segmenta-
tion. As a result, only preliminary attempts at annotating texts according to Grosz and
Sidner's theory have been made.

    For this reason, most previous corpus-based studies of centering either ignored
segmentation or used heuristics such as those proposed by Walker (1989): Consider
every paragraph as a separate discourse segment, except when its first sentence con-
tains a pronoun in subject position or a pronoun whose agreement features are not

---

18 In previous work (Poesio and Vieira 1998) we came to the conclusion that kappa, while appropriate
   when the number of categories is fixed and relatively small, is problematic for anaphoric reference,
   when neither condition applies, and may result in inflated values of agreement.

matched by any other CF in the same sentence. We tested only heuristic methods as well, using the layout structure of the texts as a rough indicator of discourse structure. In this article we discuss only the results with the heuristic proposed by Walker. In the extended technical report available on the Web site, we discuss the results with other segmentation heuristics, as well as further results with the tutorial dialogues subdomain of the GNOME corpus, independently annotated according to relational discourse analysis (Moser and Moore 1996), a technique inspired by Grosz and Sidner's proposals, from which a Grosz and Sidner–like segmentation was extracted as proposed in Poesio and Di Eugenio (2001).

3.4 Automatic Computation of Centering Information
The annotated corpus is used by Perl scripts that automatically compute the centering data structures (utterances, CFs, and CBs) according to the particular parameter instantiation chosen, find violations of Constraint 1, Rule 1, and Rule 2 (according to several versions of Rule 1 and Rule 2), and evaluate the claims using the statistical tests. The behavior of the scripts is controlled by a number of parameters, including

CBdef: which definition of CB should be used. (All the results discussed in this article were computed using the definition in Constraint 3.)

uttdef: identify utterances with sentences, finite clauses, or verbed clauses.

previous utterance: treat adjunct clauses Kameyama-style or Suri and McCoy-style.

realization: allow only direct realization, or allow indirect realization as well.

CF-filter: treat all NPs as introducing CFs, or exclude certain classes. At the moment it is possible to omit first- and second-person NPs and/or NPs in predicative position (e.g., *a policeman* in *John is a policeman*).

rank: rank CFs according to grammatical function, linear order, a combination of the two as in Gordon, Grosz, and Gillion (1993), or information status as in Strube and Hahn (1999).

prodef: consider as R1-pronouns only third-person personal pronouns (*it*, *they*), or include also demonstrative pronouns (*that*, *these*), and/or the second-person pronoun (*you*).

segmentation: identify segments using Walker's heuristics, or with paragraphs, sections, or whole texts.

Among the many other script parameters whose effect will not be discussed here we will just mention those that determine whether implicit anaphors in bridging references should be treated as CFs, the relative ranking of entities in complex NPs, and how to handle "preposed" adjunct clauses. (See the extended technical report on the Web site.) The algorithm used to compute statistics concerning violations of the claims is fairly straightforward, and we will therefore omit it here; the interested reader can find a discussion in the extended technical report. The one additional complication that we need to mention here is relative pronouns. As it could be argued that the decision to generate a relative pronoun is primarily controlled by grammatical considerations, we attempted to ignore them as much as possible, in the following sense. Our scripts do not count an utterance as a violation/verification of Rule 1 from Grosz, Joshi, and Weinstein (1995) if the only "pronoun" realizing a non-CB is a relative pronoun, or the CB is realized only by a relative pronoun. What this means in practice is that the

Table 1
Number of utterances and CFs with the vanilla instantiation.

	Museum	Pharmaceutical	Total
Number of utterances	430	577	1,007
Number of utterances that are segment boundaries	91	134	225
Number of CFs	1,731	1,308	3,039

number of utterances examined to evaluate Rule 1 is generally less than the number of utterances with a CB, as we will see shortly.

4. Main Results

Given the number of parameters, it is difficult, if not impossible, to discuss the results with all instantiations. Instead, we begin by discussing the results with what we call the "vanilla instantiation," based on the settings for the parameters most often used in discussions of the theory. We then examine the results obtained by varying the definitions of utterance and realization. After establishing the "best" values for these parameters, we consider the effect of alternative ranking functions. Additional results are discussed in the extended technical report available on the companion Web site. Readers who want to try out instantiations not discussed here should visit the Web site.

4.1 The Vanilla Instantiation
What we call the "vanilla instantiation" is not an instantiation actually proposed in the literature, but an attempt to come as close as possible to a "mainstream" instantiation of centering by blending proposals from Grosz, Joshi, and Weinstein (1995) and Brennan, Friedman, and Pollard (1987) and incorporating additional suggestions from Kameyama (1998) and Walker, Joshi, and Prince (1998a). The vanilla instantiation is based on the definition of CB from Grosz, Joshi, and Weinstein (1995) and uses grammatical function for ranking, as proposed there and in Brennan, Friedman, Pollard (1987). Because Grosz et al. do not provide a definition of utterance, the vanilla instantiation incorporates the hypothesis from Kameyama (1998) that utterances are finite clauses and the characterization of "previous utterance" proposed there.[19] Concerning realization, in the vanilla instantiation only third-person NPs introduce CFs, and a discourse entity counts as "realized" in an utterance only if it is explicitly mentioned. For the purposes of Rule 1, we mainly studied a "strict" definition of R1-pronoun including only personal (and possessive) pronouns and relative pronouns and traces (see Walker, Joshi, and Prince [1998a], page 4), but we also considered a "broader" definition including the demonstrative pronouns *this*, *that*, *these*, and *those*. Relative clauses are assumed to include a link to the NP they modify, possibly not explicitly realized. The segmentation heuristic proposed by Walker (1989) is adopted. With the parameters set in this way, the number of utterances and CFs in our corpus is as shown in Table 1.

---

19 We simplified Kameyama's hypothesis about relative clauses by considering only instantiations in which they were treated as utterances both "locally" and "globally", and ones in which they weren't.

Table 2
Utterances and CBs with the vanilla instantiation.

	Museum	Pharmaceutical	Total Percentage
Number of times at least one CF(U_n) is realized in U_{n+1}	195	162	357 (35.4%)
Utterances that satisfy Constraint 1 (have exactly one CB)	189	157	346 (34.4%)
Utterances with more than one CB	6	5	11 (1.1%)
Utterances without a CB but are segment boundary	67	96	163 (16.2%)
Utterances without a CB	168	319	487 (48.4%)

4.1.1 Constraint 1. The statistics relevant to Constraint 1 (that utterances have exactly one/at most one CB) are shown in Table 2. This table clearly indicates that the weak version of Constraint 1 (Weak C1) is likely to be verified with the vanilla instantiation. Even without counting segment boundaries, Weak C1 is verified by 833 utterances out of 1,007 (82.7%) and violated by only 11 (1.1%): The chance that Weak C1 will not hold with a different sample is $p \leq 0.001$ by the sign test. (We will henceforth write +833, −11 to indicate numbers of verifiers and violators.) On the other hand, the strong version of C1—that every utterance has exactly one CB—is not likely to hold with this instantiation: In our corpus, only 346 utterances (34.4%) have exactly one CB, whereas 498 utterances (49.4%) have zero or more than one CB. With +346, −498, the chance of error in rejecting the null hypothesis that Strong C1 doesn't hold is obviously much higher than 10%. The chance of error doesn't go below 10% even if we count the 163 utterances that do not contain references to CFs introduced in the previous utterance but are segment boundaries and therefore are not governed by the constraint. In other words, if the vanilla instantiation were the "right" way of setting the parameters, we would have to conclude that in the genres contained in our corpus, utterances are very likely to have a unique CB, but entity coherence does not play a major role in ensuring that a text is coherent: only 35.4% of utterances in our corpus would be "entity-coherent," that is, would contain an explicit mention to an entity realized in the previous finite clause.

The following example illustrates why there are so many violations of Strong C1 with the vanilla instantiation. If we identify utterances with finite clauses, the two sentences in (9) break up into five utterances, and only the last of these can be considered in any sense to directly refer to the set of egg vases introduced in (u1).[20]

(9)   (u1) These "egg vases" are of exceptional quality: (u2) basketwork bases
       support egg-shaped bodies (u3) and bundles of straw form the handles,
       (u4) while small eggs resting in straw nests serve as the finial for
       each lid. (u5) Each vase is decorated with inlaid decoration: . . .

Clearly, there are two ways of "fixing" this problem. One is to claim that utterances are best identified with sentences, in which case we would have only two utterances in this example, one for each sentence. The other is to allow for indirect realization:

---

20 In fact, the anaphoric relation here is not identity; rather, the set of egg vases serves as domain restriction for the quantifier in (u5). We were not able to mark this distinction reliably.

(u2)–(u4) all contain implicit references to the egg vases, and therefore all will have a CB if indirect realization is allowed. Both possibilities are considered below.

The fact that 11 utterances (1.1%) in the corpus have *more than one* CB (i.e., they violate Weak C1 as well) is also worth noticing. The reason for the violation is that in "classic" centering, ranking is only required to be a partial order (see, e.g., Walker, Joshi, and Prince 1998a page 3),[21] so when two CFs with the same rank in U_i are both realized in U_{i+1}, both become the CB. This is illustrated in (10), where we show the XML markup so that the attributes of elements are visible:

```
(10)    <unit finite='finite-yes' id='u227'>
            <ne id='ne546' gf='subj'>The drawing of
                <ne id='ne547' gf='np-compl'>the corner cupboard</ne></ne>
            <unit finite='no-finite' id='u228'>,   or more probably
                <ne id='ne548' gf='no-gf'> an engraving of
                    <ne id='ne549' gf='np-compl'>  it </ne></ne>
            </unit>,
            must have caught
            <ne id='ne550' gf='obj'>
                <ne id='ne551' gf='gen'>Branicki's </ne> attention</ne>
        </unit>
        <unit id="u229" finite="finite-yes">
            <ne gf="subj" id="ne552">Dubois</ne> was commissioned through
                <ne gf="adjunct" id="ne553">  a Warsaw dealer </ne>
                <unit id="u230" finite="finite-no"> to construct
                    <ne gf="obj" id="ne554">  the cabinet </ne>
                  for<ne gf="adjunct" id="ne555">the Polish aristocrat</ne>
                </unit>
        </unit>
        <ante current='ne554' rel='ident'><anchor antecedent='ne549'>
        </anchor></ante>
        <ante current='ne555' rel='ident'><anchor antecedent='ne551'>
        </anchor></ante>
```

In this example, two discourse entities introduced in utterance (u227) are realized in utterance (u229):[22] *the corner cupboard* (realized in (u227) by (ne547) and (ne549) and in (u229) by (ne554)) and *Branicki* (realized in (u227) by (ne551), and in (u229) by (ne555)). As their grammatical functions are equivalent under the ranking proposed by Grosz et al., (np-compl, for NP-complement, and gen, for genitive—see the annotation manual available on the Web site), these two CFs have the same rank in (u227), so they are both CBs of (u229). The same problem occurs with coordinated NPs, both of which have the same grammatical function. This problem with the vanilla instantiation can also be "fixed" by requiring the ranking function to be a total order, which is easily done by adding a disambiguation factor such as linear order, as done by Strube and Hahn. On the other hand, the requirement that ranking be total has not been previously discussed in the centering literature, and one might argue conversely that examples such as the one above are arguments againts centering's claim that utterances have only one CB. We return to this issue in Section 5.

---

21 It's not clear to us why ranking is required only to be partial, yet the CB is clearly claimed to be unique.
22 Neither (u228) nor (u230) is treated as an utterance as they are not finite.

Table 3
Transition statistics for the Brennan et al. version of Rule 2.

	Museum	Pharmaceutical	Total (Percentages)
Establishment	96	93	189 (18.8%)
Continuation	37	33	70 (7.0%)
Retain	22	16	38 (3.8%)
Smooth Shift	22	15	37 (3.7%)
Rough Shift	18	5	23 (2.3%)
ZERO	87	81	168 (16.7%)
NULL	148	334	482 (47.9%)
Total	430	577	1,007

4.1.2 Rule 2. The statistics relevant for Brennan et al.'s version of Rule 2 are shown in Table 3. The most obvious consideration suggested by this table is that the three most frequent transitions in our corpus are ones that either have not been previously discussed in the Centering literature or have been discussed only in a limited way. By far the most frequent transition (47.9% of the total) is NULL: following up an utterance without a CB with a second one, also without a CB. (Examples include (u3), (u4), and (u5) in (9).) We found this transition discussed only in Passonneau (1998). The second most common transition (18.8%) is Kameyama's Establishment (the transition between an utterance without a CB and one with a CB), followed by its reverse, the ZERO transition (between an utterance with a CB and one without), never mentioned in the literature. (An example of ZERO is (u2) in (9).) If we ignore NULL, Establishments, and ZEROs, the preferences are roughly as predicted by Brennan et al.: The Page test for ordered alternatives (Siegal and Castellan 1988, pages 184–188) indicates a chance of less than .001 that the four transitions are equally likely. But only the differences between CON and RET/SSH, and between SSH and RSH, are significant; and there are more Shifts (SSH + RSH) than Retains.[23]

Grosz et al.'s formulation of Rule 2 in terms of sequences also roughly holds, except that there are too few sequences for the results to be considered significant, as shown in Table 4. As we'll see again in Section 5, in our corpus there seems to be a preference for avoiding repetition; this tendency is confirmed by the numbers in the table, which indicate a dispreference for maintaining the same CB for too long or for maintaining it in the most salient position, at least at the level of finite clauses: EST/CON sequences are twice as common as sequences of Continuations. As for the claim that Retaining transitions prepare for Shifts, the figures do not lend much support to the idea: Retains are more frequently followed by Continuations than by Shifts, and almost as frequently by other Retains.

Of the other formulations of Rule 2, the version based on a preference for cheap transition pairs over expensive ones proposed by Strube and Hahn is not verified with the ranking function used in the vanilla instantiation—which is not, we should em-

---

23 Similar results were obtained by Passonneau (1998).

Table 4
Rule 2 statistics considering sequences of transitions.

	Museum	Pharmaceutical	Total
Continuation Sequences	10	6	16
Continuation/Retain	9	3	12
Establishment/Continuation	17	18	35
Retain Sequences	15	3	8
Retain/Continuation	7	7	14
Retain/Smooth Shift	3	2	5
Retain/Rough Shift	4	1	5
Smooth Shift Sequences	2	1	3
Rough Shift Sequences	2	1	3
Null Sequences	95	228	323
Other	290	312	602

Table 5
Cheap and expensive transitions with the vanilla instantiation.

	Museum	Pharmaceutical	Total
Cheap transitions	76	63	139
Expensive transitions	263	380	643
Cheap transition pairs	21	14	35
Expensive transition pairs	162	234	396

phasize, the one assumed by Strube and Hahn themselves.[24] Ignoring the 225 segment boundary utterances, we find 396 pairs of expensive transitions and 35 pairs of cheap transitions, as shown in Table 5. These statistics mean that in only 139 cases out of 357 (the total number of entity-coherent utterances with this instantiation; see Table 2), $CB(U_i)$ is predicted by $CP(U_{i-1})$. We do find that 219 utterances, the majority (61.3%) of entity-coherent ones, are "salient" in the sense of Kibble (2001)—that is, their CB is the same as their CP.

4.1.3 Salience and Pronominalization. The statistics for pronominalization are shown in Table 6. As noted above, our corpus contains 217 uses of third-person pronouns, 23 demonstratives, and 78 complementizers.[25] In this instantiation we take R1-pronouns to include only personal pronouns and complementizers, for a total of 295 R1-pronouns. If we identify utterances with finite clauses, 61 personal pronouns (28.1%) have their

24 Similar results were found for dialogues by Byron and Stent (1998).
25 We will use the term *complementizer* to indicate relative pronouns and relative traces.

Table 6
R1-pronouns in the corpus with the vanilla instantiation.

	Museum	Pharmaceutical	Total
Total number of R1-pronouns	200	95	295
Number of personal pronouns	144	73	217
Number of complementizers	56	22	78

Table 7
CBs and pronominalization with the vanilla instantiation.

	Museum	Pharmaceuticla	Total (Percentage)
Total number of realizations of CBs	211	163	374
Total number of CBs realized as R1-pronouns	138	68	206 (55.1%)
CBs realized as personal pronouns	85	48	133 (35.6%)
CBs realized as complementizers	53	20	73 (19.5%)
CBs *not* realized as R1-pronouns	73	95	168 (44.9%)
Total number of R1-pronouns that do not realize CBs	58	23	81 (27.5%)
Personal pronouns that do not realize CBs	55	21	76 (35.0%)
Complementizers that do not realize CBs	3	2	5 (6.4%)

antecedent in the same utterance, and 28 (13%) are "long-distance pronouns" (Hitze-man and Poesio 1998) whose antecedent is in neither the same nor the previous utter-ance.

Table 7 shows that the relation between pronominalization and CB with the vanilla instantiation is not straightforward: Only 55.1% of the 374 mentions of CBs[26] are pronominalized. And if relative clause complementizers were not included among the R1-pronouns (on the grounds that the decision to use a complementizer is primar-ily dictated by grammatical, rather than discourse, considerations), more CBs would be realized as non-R1 pronouns (171, 44.9%) than as R1-pronouns (137, 35.9%). On the other hand, 73% of R1-pronouns do refer to the CB.[27]

Table 8 analyzes pronominalization in terms of the three versions of Rule 1 we are considering.[28] Given the figures in Table 7, it should already be clear that the stronger version of Rule 1 we considered, always pronominalize the CB—generalizing the proposal by Gordon, Grosz, and Gillion (1993) to the less restrictive definition of CB given by Constraint 3—is not verified: In fact, 55% of utterances violate it. The two other versions of Rule 1 we are considering, however—Rule 1 (GJW 83), pronominalize

---

26 Even though only 357 utterances have a CB with this instantiation, a CB may be realized more than once in an utterance.
27 Earlier versions of these findings led to the development of the pronominalization algorithm in Henschel, Cheng, and Poesio (2000).
28 As discussed in Section 3, what is counted here are utterances that verify or violate Rule 1. Not all utterances are considered: of the 346 utterances that have exactly one CB, 72 are ignored by the script in that the only realization of an R1-pronoun is done via a relative pronoun or trace, so only 274 (27.2% of the total number of utterances) are considered relevant for Rule 1.

Table 8
Evaluation of the different versions of Rule 1 with the vanilla instantiation.

	Museum	Pharmaceutical	Total (Percentage)
GJW 95–utterances that satisfy	130	135	265 (96.7%)
GJW 95—utterances that violate	7	2	9 (3.3%)
GJW 83—utterances that satisfy	117	105	222 (81%)
GJW 83—utterances that violate	20	32	52 (19.0%)
Gordon—utterances that satisfy	77	45	122 (44.5%)
Gordon—utterances that violate	60	92	152 (55.5%)

the CB if it's the same as the CB of the previous utterance, and especially Rule 1 (GJW 95), pronominalize the CB if anything else is—are verified by most utterances.

There are two classes of violations of Rule 1 (GJW 95): possessive pronouns and pronouns referring to "global topics". In (11), CB(u3), *PRODUCT-X*, is realized as a proper noun, whereas a possessive pronoun is used to refer intrasententially to *the baby*.[29]

(11)   (u1) Infants and children must not be treated continuously with
         PRODUCT-X for long periods
         (u2) because it may reduce the activity of the adrenal glands, and so
         lower resistance to disease.
         (u3) Similar effects on a baby may occur after extensive use of
         PRODUCT-X by its mother during the last weeks of pregnancy
         (u4) or when she is breastfeeding the baby.

In the pharmaceutical leaflets, several violations of Rule 1 are found toward the end of text, when pronouns are sometimes used to realize the product described by the leaflet. For example, *it* in the following example refers to the cream discussed by the leaflet, not mentioned in the previous two utterances.

(12)   (u1) A child of 4 years needs about a third of the adult amount. (u2) A
         course of treatment for a child should not normally last more than five
         days (u3) unless your doctor has told you to use <u>it</u> for longer.

What we seem to observe here is a conflict between the "global" preference to realize the "main character" of a discourse as a pronoun, and the "local" preference to pronominalize the locally most salient entity, as identified by the CB.[30] By the end of a leaflet, the product discussed in the leaflet has been mentioned a number of times, so that it is salient enough to justify pronominalization even when it is not in CF list.

We saw in Table 7 that although there are only 9 violations of Rule 1 from Grosz, Joshi, and Weinstein (1995), 81 R1-pronouns do not realize CBs. The majority of the

---

29 The problem of intrasentential pronouns in Centering is discussed, e.g., in Walker (1989), Tetreault (2001), and Poesio and Stevenson (forthcoming).
30 See also Giouli (1996) and Byron and Stent (1998).

72 cases of pronouns that do not refer to the CB but do not violate Rule 1 fall into two classes: (1) R1-pronouns used in utterances without a CB (the majority) and (2) R1-pronouns used in utterances in which the CB is pronominalized as well—as in the following example, in which both *the microscope* and *the amateur scientist* are realized (by a personal pronoun and a relative trace) in the relative clause (u2):

(13)   (u1) This microscope belonged to an amateur scientist,
         (u2) who would have used it to explore the mysteries of the natural
         world.

82.6% of demonstrative pronouns in the corpus do not realize the CB, which is what one would expect on the basis of, for example, Gundel, Hedberg, and Zacharski (1993), and Passonneau (1993). This suggests that treating demonstrative pronouns as R1-pronouns would not lead to improvements with respect to Rule 1. On the other hand, because there are only 23 demonstrative pronouns in the corpus, such a change would be unlikely to drastically affect the results. And indeed, with a broader definition of R1-pronoun that includes demonstrative pronouns, we find a few more violations of Rule 1 (GJW 95) (11 instead of 9) and a few less violations of Rule 1 (Gordon et al.) (148 instead of 152) and of Rule 1 (GJW 83) (50 instead of 52), but none of these differences is significant. The results reported in the rest of the article are all obtained with the "narrow" definition of R1-pronoun that does not include demonstratives.[31]

4.1.4 Differences Among the Domains. The texts in the museum domain seem to be more in agreement with the predictions of the theory than the texts in the pharmaceutical domain. This is especially the case for Rule 1. There are fewer personal pronouns in the texts in the pharmaceutical domain (73 of 1,308 CFs, [5%], as opposed to 144 of 1,731 [8%] for the museum domain), and whereas in the museum domain 40.3% (85/211) of CB realizations are done via personal pronouns (65.4% if we consider all R1-pronouns), in the pharmaceutical domain only 29.4% (48/163) are (41.7% for R1-pronouns). The percentage of utterances satisfying the strong version of Constraint 1 is much higher in the museum domain (44.0%, 189/430) than in the pharmaceutical one (27.2%, 157/577), and the percentage of utterances with no CB that are not segment boundaries is much higher in this second domain (55.3%, 319/577) than in the first (39.1%, 168/430). Finally, almost 72% of utterances in the pharmaceutical domain are NULL or ZERO transitions (415/577), whereas just 54.6% are in the museum domain (235/430); the percentage of EST and CON is also higher in the museum domain (133/430, 31%, versus 126/577 [21.8%]). These differences are in part due to the large number of second-person pronouns *you* in the pharmaceutical domain, so that the statistics for Constraint 1 improve if we treat the entities referred to by these pronouns as CFs, as we will see below. A second reason for these differences is that layout plays a much more important role in the pharmaceutical domain, providing a different way of achieving coherence. (See Section 5.)

4.2 Varying the Utterance Parameters
We now begin to explore alternative parameter settings. As always, space constraints prevent a full discussion of all the instantiations. In this article, we discuss the results with most of the variants quite briefly and analyze at some length only the instantiation

---

31 The relation between demonstrative NPs in general and the CB in our corpus is analyzed in detail in Poesio and Nygren-Modjeska (2003).

that identifies utterances with sentences; the results are summarized with graphs. The extended technical report available on the companion Web site contains a more extensive discussion of some of the variants; interested readers are also encouraged to try further instantiations on the Web site. In this subsection we consider how the definition of utterance (parameter **uttdef**) and the value of the parameter **previous utterance** affect the claims.

4.2.1 Treating Coordinated Verb Phrases as Utterances. Many researchers working on spoken dialogues or NLG assume that each element of a coordinated VP counts as a separate utterance: that is, that in *We should send the engine to Avon and hook it to the tanker car*, the coordinate VP *hook it to the tanker car* is actually a separate utterance. Treating coordinated VPs as separate utterances in our corpus would of course result in more utterances (1,041 vs. 1,007), which of course would lead to worse results unless these utterances were treated as containing an implicit trace. If we do so, we obtain slightly (but significantly) better results for Strong C1 (48% violations instead of 49%), and nonsignificant differences for Rule 1 and Rule 2 (with slightly higher numbers of Continuations and slightly lower number of Retains).

4.2.2 Using All Verbed Clauses Instead of Just the Finite Ones. A second extension of the definition of utterance is to treat as utterances *all* clauses with a verb, including, for example, the infinitival *to*-clause in *John wants to visit Bill*. The results with such an instantiation, as well, crucially depend on our grammatical assumptions. With such a definition, we get of course many more utterances (1,267 instead of 1,007), most of which, like the example infinitival clause just given, do not contain explicit mentions of the argument in subject position; so again, if we didn't assume that traces are present in such clauses, we would find significantly more violations of Strong C1 (685 instead of 498). Using a crude mechanism for tracking traces (adding a trace referring to the subject of the matrix clause to all nonfinite complement clauses), we still find a larger number of violations (598) than with the vanilla instantiation, but because the number of utterances is much greater, these violations represent a significantly lower percentage of the total (47% instead of 49%). We find no significant differences in the number of violations of Rule 1. As for Rule 2, this change results in significantly fewer NULL transitions (45.0% instead of 47.9%) and significantly more EST (22.1% instead of 18.8%) and SSH (5.6% instead of 3.7%).

4.2.3 Restricting Finite Clauses. In general, the best results for C1 are obtained by considering larger chunks of text as a single utterance, thus reducing the number of utterances. In particular, fewer violations are obtained by not considering as utterances finite clauses that occur as parentheticals, as subjects (as in *That John could do this to Mary was a big surprise to me*), and as matrix clauses with an empty subject (as in *It is likely that John will arrive tomorrow*). This merging only reduces the overall number of utterances from 1,007 to 972, but the result is a simultaneous reduction in the violations of Strong C1 from 498 to 469 (48.2%) (which is significant by the binomial-proportions test, though still not enough for Strong C1 to be verified) and an increase in the number of utterances that satisfy Rule 1 (GJW 95) to 281. The violations of Rule 1 are also reduced to 8 (2.8%) (not significant). (There are virtually no changes for Rule 2.) Because of these small improvements, in the rest of the article we always exclude these occurrences when discussing the results with finite clauses as utterances; we refer to this instantiation as "vanilla−".

4.2.4 Relative Clauses. Relative clauses turned out to be one of the most complex problems we had to face. The reader may recall that Kameyama tentatively proposes (without empirical support) that relative clauses have a "mixed" status: They should be locally treated as updating the local focus, but at the global level they should be merged with the embedding utterance. This proposal, however, seems to assume that the local focus may be updated with the content of certain utterances some time after they have been first processed, which is a rather radical change to the basic assumptions of the framework. In this study, we simply compared an instantiation in which relative clauses are treated as utterances with one in which they are not. In addition, we considered treating relative clauses as adjuncts (i.e., as not embedded) and treating them as complements (embedded).[32] The figures reported so far were obtained by treating relative clauses as utterances and as akin to adjuncts.[33] Not treating relative clauses as separate utterances results in a 6.5% reduction in the number of utterances with respect to vanilla− (908 instead of 972) and in fewer violations of Strong C1 (452 [439 utterances without a CB, 13 with two CBs] instead of 469 [457 and 12]); however, the percentage of violations is higher (49.7% vs. 48.2%). The number of violations of Rule 1 also stays the same (8 [2.7%]). From the point of view of Rule 2, a lot of relative clauses seem to function as EST, since their number goes down by almost 15%, to 17.3% (from 190 to 157); we also see a 30% reduction in SSH and an increase in NULL, to 50.6% of the total. Everything else stays the same.

In purely numerical terms, then, not treating relative clauses as separate utterances would not improve the results. Furthermore, and most important, we feel that not treating finite relative clauses as separate utterances would make it very difficult to maintain the principle that utterances are identified with finite clauses. For these reasons, in the rest of the article we will continue to count relative clauses as finite clauses.

4.2.5 Suri and McCoy's Definition of Previous Utterance. As discussed in Section 2, Suri and McCoy (1994) suggested that *after-* and *before-*clauses behave more like embedding elements (i.e., like complements) than like coordinating ones, and Cooreman and Sanford (1996) found evidence supporting this treatment for *when* clauses, as well. The **previous utterance** parameter of our script can be used to compare this proposal with Kameyama's. When this parameter is set "Kameyama-style," adjunct clauses are treated as not embedded, so that, in (14), the previous utterance for (u3) is (u2). (This was the setting used for the results discussed so far.) When the parameter is set to (generalized) Suri and McCoy, adjunct clauses are treated as embedded, so that the previous utterance for (u3) is (u1):

(14)   (u1) John woke up (u2) when Bill rang the door.
         (u3) He had forgotten the appointment.

Using Suri and McCoy's definition of previous utterance results in a small but significant reduction in the number of violations of Strong C1, in small improvements concerning R1 (GJW 95), and in small, but not significant, improvements for Rule 2. As far as Strong C1 is concerned, 20 utterances that violate Strong C1 with Kameyama's

---

32 The difference matters when the relative clause occurs at the end of an embedding clause, as in *John wanted a photograph of the man that Bill had seen entering the building at night. <u>He . . .</u>*

33 We also remind the reader that our script treats all relative clauses as containing a link referring to the entity modified by the relative, even when the clause does not contain an explicit relative pronoun or complementizer, so that they never violate C1.

definition satisfy it under Suri and McCoy's, but 9 utterances become violations (by the sign test, +20, −9, $p \leq .03$). The reduction is not, however, sufficient for Strong C1 to be verified (+355, −458). With Rule 1, we find that the number of utterances that verify the GJW 95 version increases (+287), but the number and percentage of violations stays the same (8 [about 2%]).

We should note, however, that the differences between Kameyama's and Suri's definition of previous utterance have mostly to do with a type of clause that was discussed only briefly by Kameyama and not at all by Suri and McCoy: relative clauses, as in (15):

(15)   (u1) This brooch is made of titanium,
        (u2) which is one of the refractory metals.
        (u3) It was made by Anne-Marie Shillitoe, an Edinburgh jeweller, in 1991.

If the "generalized Kameyama" definition of previous utterance is adopted, the previous utterance for (u3) is the relative clause, (u2); this causes a violation of Strong C1. In the 'generalized Suri and McCoy' instantiation, by contrast, the relative clause is treated as embedded; this seems to be the better approach. If relative clauses were not treated as separate utterances or were treated as embedded in both instantiations, we would find an equal number of violations, although about 20 violations would be different in each instantiation. One example in which the difference does have to do with the way adjuncts are handled is (7). PRODUCT-Z is not mentioned in the adjunct *if*-clause, and therefore Strong C1 is violated if (u2) is taken as previous utterance for (u3). In this case, Suri and McCoy's proposal works better than Kameyama's.

(7)   (u1) You should not use PRODUCT-Z
       (u2) if you are pregnant or breast-feeding.
       (u3) Whilst you are receiving PRODUCT-Z ....

Conversely, in the following example, the adjunct clause, *as you may damage the patch inside*, introduces the entity *the patch*, which is then referred to in (u3), so treating the adjunct (u2) as embedded leads to a violation of C1. In this case, Kameyama's definition of previous utterance gives the right result:

(16)   (u1) Do not use scissors
        (u2) as you may damage the patch inside.
        (u3) Take out the patch.

Given that these improvements are significant, if small, in the rest of the article, we will use Suri and McCoy's definition when **uttdef** is set to *finite clause*. However, our discussion, and especially the contrast between (7) and (16), gives further support to the idea that utterances may be best identified with sentences. We consider this setting next.

4.2.6 Sentences. The setting of **uttdef** with the most dramatic impact on Strong C1 is that which identifies utterances with sentences. The reasons for this have already been illustrated with (9): If utterances are identified with sentences, there are only two utterances in that example, both containing references to the egg vases. The reduction in violations is such that with this instantiation more utterances verify Strong C1

Table 9
Statistics relevant to Constraint 1 when utterances are identified with sentences.

	Museum	Pharmaceutical	Total (Percentage)
Number of times at least one CF(Un) is realized in Un+1	131	147	278 (41.6%)
Utterances that satisfy Constraint 1 (have exactly one CB)	126	138	264 (39.5%)
Utterances with more than one CB	5	9	14 (2.1%)
Utterances without a CB but segment boundary	65	80	145 (21.7%)
Utterances without a CB	75	171	246 (36.8%)

than violate it, although not so many as to ensure verification at the 5% level.[34] The statistics relevant to Constraint 1 with this definition of utterance are shown in Table 9. Although Strong C1 is still not verified if we consider all 669 segments of text that contain NPs, the number of utterances that satisfy Strong C1 (264) is slightly larger than the number of those that don't (260).

However, identifying utterances with sentences also has several negative (if small) effects. The main among these is that the number of violations of Rule 1 goes up: In the case of Rule 1 (GJW 95), by 50%, from 8 to 12. The reason for this increase is in part simply that more utterances have a CB; but in some cases, the problem could be viewed as the CB's not being updated quickly enough. Consider the following example:

(17)   (s1) The engravings for these rooms, showing the wall lights in place, were reproduced in Diderot's Encyclopedie, one of the principal works of the Age of Enlightenment. (s2) An inscription on the Getty Museum's drawing for one of these wall lights explains (cl3) that <u>it</u> should hang above the fireplace.

The pronoun *it* in sentence (s2) violates Rule 1 if utterances are viewed as sentences, but not if they are viewed as clauses. This is because in the first case (s2) has a single CB, *the wall lights*, whereas with the vanilla− instantiation clause, (cl3) is a separate utterance, with CB *one of these wall lights*. Because the number of violations is still quite small, both Rule 1 (GJW 95) and Rule 1 (GJW 83) are still verified (+252, −12; and +209, −55, respectively, as opposed to +287, −8 and +243, −52, with the vanilla− instantiation, Suri setting),[35] although Rule 1 (Gordon et al.) still isn't (+97, −167). Note also that with this instantiation, the number of CBs realized by R1-pronouns (129) is much smaller than the number realized by other types of NPs (209).

---

34 There is one complication: Many CFs are introduced not in sentences, but in in titles and other layout elements that do not have a sentential format, such as *Chandelier* or *Side effects*. In order not to leave these CFs "stranded," the scripts also treat as an utterance every unit that contains an NP which is not contained in any sentence, just as we did for the vanilla instantiation. This means, however, that the number of utterances is much larger than the number of sentences (669 instead of 505), and that Strong C1 is not verified, even though it would be if only the 505 sentences were considered (the sign test would then show $p \leq 0.001$ for Strong C1).

35 The number of utterances to be tested of course varies depending on whether utterances are identified with finite clauses (295) or sentences (264).

Table 10
Rule 2 statistics with sentences as utterances.

	Museum	Pharmaceutical	Total (Percentage)
Establishments	54	68	122 (18.2%)
Continuations	28	33	61 (9.1%)
Retain	22	23	45 (6.7%)
Smooth Shift	7	12	19 (2.8%)
Rough Shift	20	11	31 (4.6%)
ZERO	52	66	118 (17.6%)
NULL	88	185	273 (40.8%)

The results for Rule 2 are not that different from those obtained with finite clauses, but we do observe more Continuations and fewer NULLs. The statistics are shown in Table 10. Note the much greater number of Rough Shifts than of Smooth Shifts, although the ranking suggested by Brennan et al. is still verified by the Page test.

There are still too few sequences to truly test the version of Rule 2 proposed by Grosz et al., but the preferences are roughly verified. As for the version of Rule 2 proposed by Strube and Hahn, there still 10 times as many expensive-expensive sequences (191) as cheap-cheap ones (18).

4.2.7 Interim Summary. The effect of the changes in the definition of utterance and previous utterance on Strong C1 and Rule 1 are summarized in Figure 1 and Figure 2, respectively. As the figures show, most such changes have fairly small effects, even though the effects are often significant. The one exception is identifying utterances with sentences; treating all clauses as utterances also has a positive impact, provided that we assume nonfinite clauses contain an implicit realization of the subject of the matrix clause.

Even though identifying utterances with sentences leads to much better results for Strong C1, we will not simply abandon the hypothesis that utterances may coincide with finite clauses. This is in part for theoretical reasons, such as the fact that in other theories of discourse in which "units" are assumed, such as RST, these units are generally finite clauses. Also, identifying utterances with sentences leads to small, but significant, increases in the number of violations of Rule 1 (from 8 in the vanilla instantiation [2.8%] to 12 [4.5%]) and in the number of Rough Shifts (from 2.9% to 4.6%). We will also see in a moment that there are other ways of changing the vanilla instantiation that satisfy Strong C1, so identifying utterances with sentences is not strictly necessary.

In the rest of the article we will, therefore, study the effect of changes to other parameters both on instantiations in which utterances are identified with finite clauses (henceforth, $u = f$) and on instantiations in which they are identified with sentences ($u = s$).

4.3 Realization
In this section we discuss the effect of changes in the values of the realization parameters: **realization** and **CF-filter**.
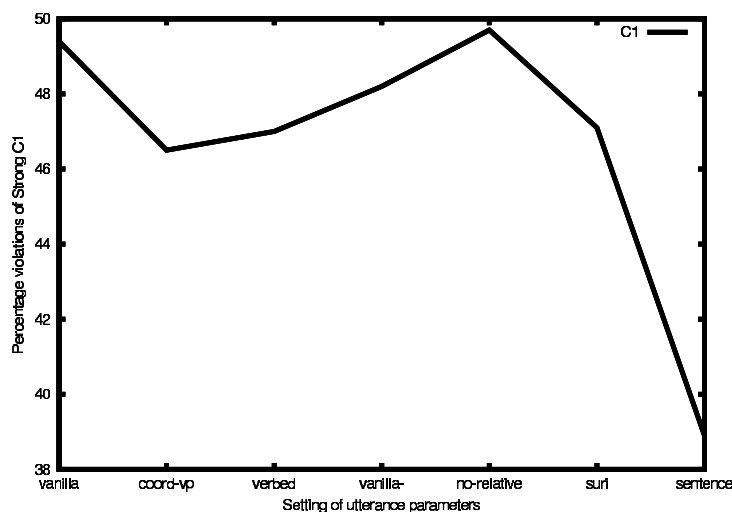
Figure 1
The effect of utterance parameters on Strong C1: A summary.
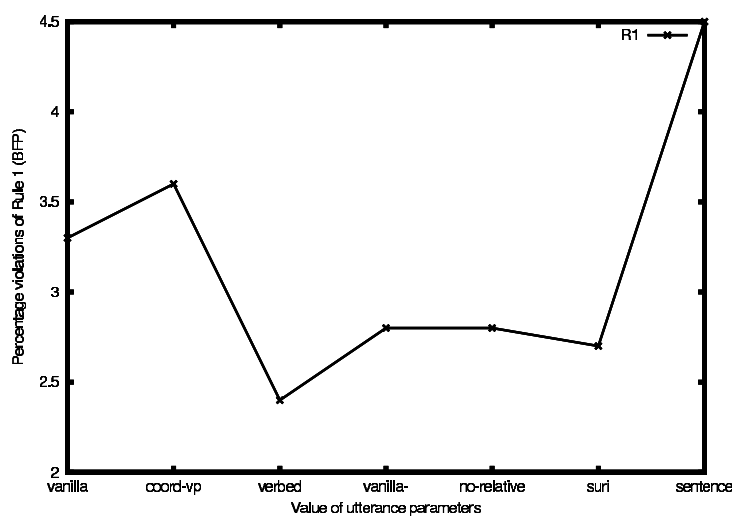
Figure 2
The effect of utterance parameters on Rule 1 (GJW 95): A summary.

4.3.1 IF: Indirect Realization + $u = f$**.** Examples such as (9) indicate that another way to reduce the number of violations of Strong Constraint 1 is to allow for indirect realization: Then the bridging references to the egg vases in (u2), (u3), and (u4) would make them the CB of these utterances. And indeed, if we modify the "best" among the $u = f$ instantiations (vanilla−, using our generalization of Suri and McCoy's proposal to determine the previous utterance) to allow for indirect realization, the reduction in violations to Strong C1 is such that with 525 utterances (54.0%) having exactly one CB and 325 having zero or more than one (33.4%), Strong C1 is verified by the sign test (+525, −325).[36]

---

36 The number of utterances is obviously not affected by changes in the realization parameters.

Table 11
Rule 2 statistics with indirect realization, $4u = f$.

	Museum	Pharmaceutical	Total (Percentage)
Establishments	75	95	170 (17.5%)
Continuations	49	40	89 (9.2%)
Retain	76	51	127 (13.1%)
Smooth Shift	39	25	64 (6.6%)
Rough Shift	60	37	97 (10%)
ZERO	60	78	138 (14.2%)
NULL	46	241	287 (29.5%)

However, allowing for indirect realization has a negative effect on other claims, just as the change to u=s does. The first negative effect is that the number of utterances with more than one CB almost doubles, from 13 with the "generalized Suri and McCoy" instantiation (1.3%) to 22 (2.3%). This is because by increasing the number of "persistent entities," we increase the chance of their having an equivalent ranking in the previous utterance. The number of violations of Rule 1 exactly doubles: from 8 with the Suri and McCoy instantiation to 16. But because with indirect realization more utterances have a CB, the number of utterances that matter for the purposes of Rule 1 also increases, from 295 to 467, so that the percentage of violations to Rule 1 does not change that much. With indirect realization, 3.4% of utterances violate Rule 1 (GJW 95), as opposed to 2.7% with generalized Suri and direct realization. As a result, Rule 1 (GJW 95) (+451, −16) and Rule 1 (GJW 83) (+318, −149) are still verified, whereas Rule 1 (Gordon et al.) still isn't (+136, −331). An example of a pronoun that becomes a violation of Rule 1 (GJW 95) if we allow CFs to be indirectly realized is shown in (18). The NP *one stand* in (u42) realizes a bridging reference to the discourse entity introduced by the NP *the two stands* in (u39), which is therefore realized in (u42), and thus becomes its CB, but it is not pronominalized.

(18)   (u39) *The two stands* are of the same date as the coffers, but were
         originally designed to hold rectangular cabinets.
         (u42) <u>*One stand*</u> was adapted in the late 1700s or early 1800s century to
         make <u>it</u> the same height as *the other*.

Finally, the change to indirect realization has a big impact on the statistics for Rule 2, shown in Table 11. On the positive side, the number of NULL transitions goes down significantly (to less than 30%), and the percentages of the four "classic" transitions go up. However, the greatest increases are in the number of RET (from 3.8% to 13.1%) and RSH (from 2.6% to 10.0%). The fact that there are many more RET than CON and many more RSH than SSH means that this is the first instantiation for which Rule 2 (BFP) is *not* verified by a Page test. The reason for this can be seen in (18): Because implicit realizations are implicit NP modifiers (i.e., *one stand* is interpreted as *one of the two stands*), they are never CPs of an utterance. (Rule 2 [Strube and Hahn] still isn't verified, although the percentage of cheap transitions increases from 154/747, [20.6%] to 207/747 [27.7%]).

Table 12
Statistics about Strong C1 with $u = s$ and indirect realization

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(U_n) is realized in $U_n + 1$	194	222	416 (62.2%)
Utterances that satisfy Constraint 1 (have exactly one CB)	184	206	390 (58.3%)
Utterances with more than one CB	10	16	26 (3.9%)
Utterances without a CB that are segment boundaries	47	55	102 (15.2%)
Utterances without CB	30	121	151 (22.6%)

Below, we indicate the instantiations with u=f, Suri and McCoy-style treatment of adjuncts, and direct realization as DF and those with the same settings, but indirect realization, as IF.

4.3.2 IS: Indirect Realization $+u = s$. As one might expect, the results for Constraint 1 get even better if indirect realization is combined with the u=s setting. As shown in Table 12, With this instantiation (henceforth, IS) 390 utterances out of 669 (58.3%) satisfy Strong C1, and 177 (26.5%) violate it—significantly better than the instantiation with u=s and direct realization (henceforth, DS). On the other hand, the number of utterances with more than one CB almost doubles again (and with respect to the DS instantiation), to 26 (3.9%) from 14 (2.1%).

The number of violations of Rule 1 (GJW 95), as well, doubles again with respect to the DS instantiation, from 12 (4.5%) to 26 (6.7% of the 390 utterances with a CB and an R1-pronoun). While this number of violations isn't enough to invalidate Rule 1 (GJW 95) (+364, −26), it is three times the number of violations with the vanilla instantiation. As for what we called Rule 1 (Gordon et al.), even with this instantiation more than 75% of utterances violate it: +97, −293.[37]

The results with Rule 2 are comparable to those obtained with the IF instantiation. Just as in that case, Rule 2 (BFP) is not verified according to a Page test, even though there is a great reduction in the number of NULL transitions (to 23.2%). The percentage of RET is even greater than with IF (114 [17.0%], almost twice the percentage of CON [9.4%]) as is that of Rough Shifts (100 [14.9%], almost three times the percentage of Smooth Shifts [5.2%]). If we ignore segment boundaries, cheap transitions are 136/444, 30.6% of the total (as opposed to 22.0% with DS and 27.7% with IF).

4.3.3 Second-Person CFs. Second-person pronouns (henceforth, PRO2s) are generally assumed to be used deictically rather than anaphorically (see, e.g., Di Eugenio 1998).

---

37 Some readers might think that the additional violations of Rule 1 obtained with instantiations IF and IS (such as the one in example (18)) shouldn't really count as violations of Rule 1, because bridging references such as *one stand* contain an implicit reference to *the two stands*, that is, are semantically equivalent to *one of the two stands*, and it is these implicit anaphors that satisfy Rule 1. One of the parameters not discussed here, **bridges_policy**, controls whether these implicit anaphoric references are treated as R1-pronouns. It turns out that doing this actually results in *more* violations of Rule 1, because most bridging references do not refer either to the CB of the present utterance or the CB of the previous one (see Poesio 2003), and every bridging reference not referring to the CB may cause a violation. In fact, treating these implicit anaphoric references as R1-pronouns (hence, as CFs) also dramatically increases the number of utterances with more than one CB, as well as the number of Rough Shifts. For details, see the extended report and the companion Web site.

However, it has been suggested in recent work that especially in dialogue, they may actually realize CFs (Byron and Stent 1998).[38] In our corpus, and especially in the pharmaceutical domain, PRO2s are very numerous and often seem to play an important role in maintaining the coherence of the discourse. And in fact, allowing PRO2s to count as realizations of CFs does reduce the number of violations of Strong C1 with both the $u = f$ and the $u = s$ instantiations of the theory, both with direct and with indirect realization. Even with DF (and the Suri-McCoy setting of the **previous utterance** parameter), allowing PRO2s to count as CFs is sufficient on its own to verify Strong C1: The museum domain is not affected, but in the pharmaceutical domain, the number of utterances that satisfy Strong C1 increases from 164 to 273, so that in total 464 utterances satisfy C1 and 367 violate it, which makes the constraint verified (by the sign test, $p \leq .03$). With DS, if we treat PRO2s as CFs, 332 utterances verify Strong C1, and 214 don't (as opposed to +264, −260 when PRO2s are not treated as CFs). Allowing for indirect realization we get even better results: With IF, we get +623, −242, a significant improvement even over the instantiation with direct realization and PRO2s; with IS, we get +439, −145.

The results with Rule 2 are also improved by treating PRO2s as CFs. The percentage of NULL transitions is greatly reduced (for DF, down to 35.0% [from 47.6%]; for DS, to 30.8% [from 40.8%]; for IF, to 18.2% [from 29.5%]; for IS, to 15.1% [from 23.2%]). As a result, the percentage of "continuous" transitions (Kibble 2001)—EST, CON, RET, SSH, and RSH—increases. However, RSH and SSH increase, as well as EST and CON: In the IF instantiation with PRO2s, EST are the most common transition (20.2%), but in the IS instantiation, RSH are (18.4%). Because of these increases, treating PRO2s as realizations of CFs does not fix the problem with rule 2 (BFP) observed above: the Rule still isn't verified with IF and IS.   There are no significant changes with Rule 2 (Strube and Hahn).

The results with Rule 1 crucially depend on whether we do or do not consider second-person pronouns as R1-pronouns. In either case, letting PRO2s realize CFs results in more violations of Rule 1 (GJW 95), both in absolute and in relative terms, because more utterances have a CB and therefore count as violations or verifications of the rule. But if we don't consider PRO2s as R1-pronouns, then the increase in violations is small: for DF, from 8 (2.7%) to 12 (2.9%); for DS, from 12 (4.5%) to 17 (5.1%); for IF, from 16 (3.4%) to 20 (3.5%); and for IS, from 26 (6.7%) to 31 (7.1%). If we treat PRO2s as R1-pronouns, however, the percentage of violations of Rule 1 (GJW 95) almost triples for the u=f instantiations and doubles for the u=s ones: 31 violations for DF (7.6%), 38 for DS (11.4%), 51 for IF (9.1%), and 67 for IS (15.3%). (Of course, in all of these cases, Rule 1 [GJW 95] remains verified in a statistical sense.) The reason for this increase in violations is that PRO2s do not seem to be very good indicators of the CB: About half of PRO2s are not realizations of CBs, in all instantiations. Given these results, it seems clear to us that it's not a good idea to treat PRO2s as R1-pronouns; it's less clear whether to treat them as realizing CFs. As we find the position that PRO2s play a deictic function convincing, in the rest of the article, we will not include their referents among the CFs, but we will indicate where doing so would result in major differences. The interested reader is advised to try the alternatives on the companion Web site.

4.3.4 Predicative NPs. The two alternative views considered so far about which entities to realize both result in an increase in the number of CFs. What if we were to

---

38 Walker (personal communication) also observed that in Japanese, zero pronouns—often taken as referring to the CB—are allowed to refer to second-person entities.

attempt to reduce the number of CFs instead? Prima facie, one would imagine this type of modification to have a negative impact on C1, but perhaps some of the violations of Rule 1 might disappear.

   Among the NPs that might be thought not to introduce CFs, an obvious candidate are predicative NPs, that is, NPs like *a policeman* in *John is a policeman* that play the role of predicates in the logical form of an utterance. But in fact, because our annotators were instructed to mark up *John* rather than *a policeman* as antecedent of subsequent anaphoric expression in these examples, filtering away such NPs does not have any positive result at all; on the contrary, it does have a significant (if small) negative impact on Strong C1,[39] because in some cases the annotators had been forced to mark up an NP in predicative position as the antecedent of an anaphoric expression against the instructions. Two such examples are given below. Especially in the second case, it is not clear how else the annotators could have marked the antecedent of *Bjorg*:

(19)   a.   An important artist in making these links has been <u>Yasuki Hiramatsu</u>. <u>His</u> knowledge of metalcraft allows <u>him</u> to push and play against the boundaries of what the material can physically do.
       b.   Two such jewellers are <u>Toril Bjorg</u> from Norway and Jacqueline Mina from England. It may be unsurprising that <u>Bjorg</u>, as a Scandinavian, should choose silver as her material.

In the following we will continue to treat predicative NPs as not introducing CFs.

4.3.5 Interim Summary. The realization parameters have an even greater impact than the utterance parameters, especially on Strong C1 and Rule 2. Either allowing for indirect realization or treating second-person pronouns as introducing CFs, is sufficient for Strong C1 to be verified. When the two settings are combined, a large majority of utterances verifies the constraint. On the other hand, allowing for indirect realization also results in significant increases in the number of violations of Rule 1, although overall the percentage of such violations remains pretty small. In addition, indirect realization leads to such an increase in the number of RET and RSH that there are fewer CON than RET and fewer SSH than RSH, and Rule 2 (BFP) is not verified by any of the instantiations with indirect realization we have discussed. Treating PRO2s as realizations of CFs, while sufficient to cause Strong C1 to be verified, has less of an effect on Rule 2, but when we combine this setting with the IS instantiations, we obtain an instantiation in which RSH is the most common transition. The effects of the realization choices are summarized in Figures 3 and 4.

4.4 Ranking
4.4.1 Grammatical Function + Linear Disambiguation. We observed above that because grammatical function does not always specify a unique, most highly ranked CF, when using that ranking function some utterances end up with more than one CB, which causes the violations of the weak version of Constraint 1 seen above (up to 5.7% of the total in the IF instantiation treating PRO2s as CF realizations). We also mentioned, however, that this problem can be fixed by requiring the ranking function to be a total order, which, in turn, is easily done by adding a tie-breaking factor. Given the results in Gernsbacher and Hargreaves (1988) and Gordon, Grosz, and Gillion (1993),

---

39 The difference is significantly larger for all the instantiations not treating PRO2s as CFs; larger, but not significantly so, if PRO2s are treated as CFs.
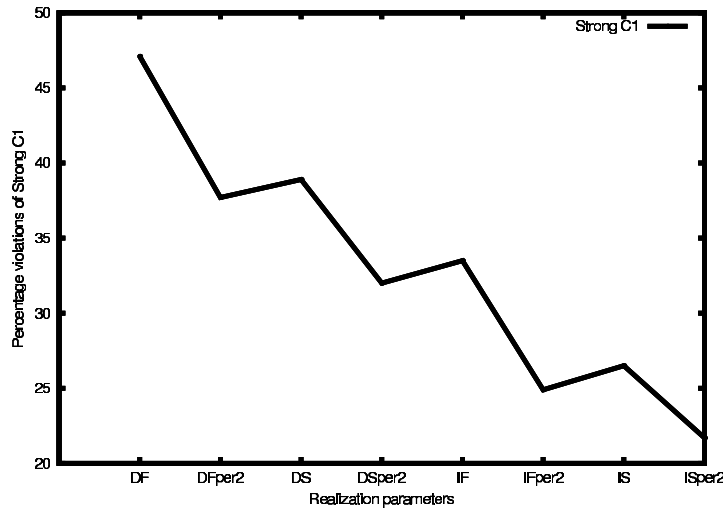
Figure 3
The effect of the realization parameters on the violations of Strong C1.
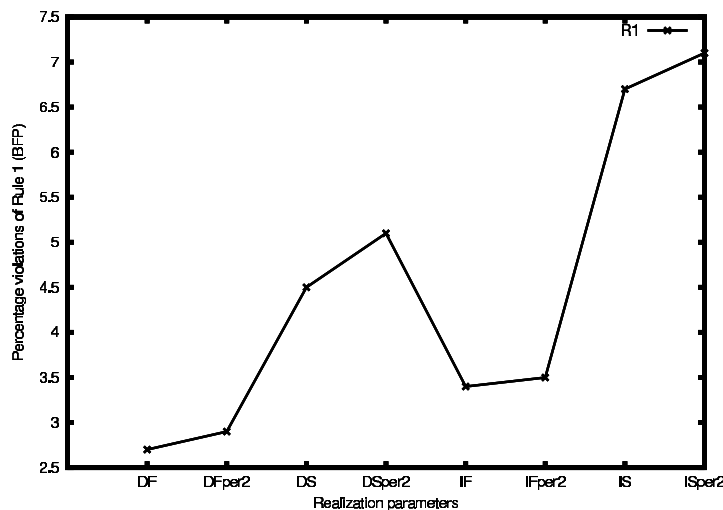
Figure 4
The effect of the realization parameters on the percentage of violations of Rule 1.

the most obvious disambiguating factor is linear order: Whenever two CFs are equally ranked, assign to, say, the leftmost CF a higher rank. And indeed, we saw in Section 2 that linear order is used by Strube and Hahn (1999) to resolve ties, albeit in conjunction with a ranking function other than grammatical function. In this section we evaluate the ranking function obtained by adding linear order to grammatical function, which we call GFTHERELIN.[40] The results with this ranking function are summarized in Table 13.

---

40 The reason for the "there" in the function's name is that the results can be slightly improved by a
   further small change: ranking postcopular NPS in *there*-sentences (e.g., *someone* in *There is someone at the door*) as subjects rather than objects. See, e.g., Sidner (1979).

Table 13
Summary of results for Strong C1, Rule 1 (GJW 95), and Rule 2 (BFP) with GFTHERELIN ranking. ♠ indicates that a claim is not verified at the .05 level; ◇ that it's not verified at the .01 level.

Instantiation	Strong C1 (Percentage violations)	Rule 1 (GJW 95) (Percentage violations)	Rule 2 (BFP) (Page test, probability of not being verified)
DF-predicate	+352,−450 (46.3%) ♠	+291,−11 (3.6%)	.001
DF-predicate+pro2	+465,−355 (36.5%)	+403,−15 (3.6%)	.001
DS-predicate	+273,−249 (37.2%) ♠	+259,−14 (5.1%)	.001
DS-predicate+pro2	+347,−197 (29.4%)	+325,−22 (6.3%)	.001
IF-predicate	+529,−310 (31.9%)	+463,−18 (3.7%)	1 ♠
IF-predicate+pro2	+635,−219 (22.5%)	+325,−22 (3.7%)	.05◇
IS-predicate	+408,−157 (23.5%)	+378,−30 (7.4%)	.05◇
IS-predicate+pro2	+469,−113 (16.9%)	+432,−37 (7.4%)	1 ♠

The table summarizes eight instantiations: DF, DS, IF, and IS, each in two variants (including PRO2s and without them). For each instantiation, the table lists verifiers and violations of Strong C1 and Rule 1 (GJW 95) and the percentage of violations, as well as the results of the Page test for Rule 2.

Adopting GFTHERELIN as a ranking function doesn't lead to major changes as far as Strong C1 is concerned. This is because the only change from the results obtained with simple grammatical function is that the utterances previously classified as having two CBs get reclassified as having one, and with the instantiations that would benefit the most from a reduction in Strong C1 violations—those based on DF—the number of multi-CB sentences is fairly small, typically 1–2%, although this is enough to make the improvement significant by the sign test with all instantiations. The improvements are greater with the u=s instantiations, since with sentences it's more common for more than one CF to be realized in the same grammatical position; for example, in the IS instantiations in which PRO2s are considered as realizations of CFs, we find that 5.7% of utterances (38/669) have more than one CB. However, Strong C1 is already verified with these instantiations, even with simple grammatical function.

As in all previous cases, better results with Strong C1 are counterbalanced by worse results for Rule 1—although, again, not so much worse that Rule 1 ends up not being verified. The results with the DF instantiations aren't significantly worse: For example, we find +291, −11 (3.6%) with the instantiation not including predicative NPs and second-person pronouns, as opposed to +280, −9 for the same instantiation with simple grammatical function ranking. The number of violations of Rule 1 is significantly greater with the DS instantiation if PRO2s are treated as CF realizations: +325, -22 (6.3%) vs. +310, -17 (5.2%). In two of the additional five violations of Rule 1, however, the problem is simply that by adding a disambiguation element, we turn utterances whose CB is undefined (because more than one CF is equally ranked) into utterances with a CB. One such example is (20):

(20)   (s7) Intended to hold jewels or small precious items, the interiors of
        this pair of <u>coffers</u> are lined with tortoiseshell and <u>brass</u> or pewter, with
        secret compartments in the base.

Table 14
Transition percentages for IS with GFTHERELIN ranking.

	Museum	Pharmaceutical	Total (Percentage)
Establishments	47	60	107 (16.0%)
Continuations	28	44	72 (10.8%)
Retain	56	65	121 (18.1%)
Smooth Shift	8	24	32 (4.8%)
Rough Shift	48	28	76 (11.4%)
ZERO	43	58	101 (15.1%)
NULL	41	119	160 (23.9%)

> (s8) <u>The coffers</u> are each decorated using techniques known as *premiere partie* marquetry, a pattern of <u>brass</u> and pewter on a tortoiseshell ground, and its reverse, *contrepartie*, a tortoiseshell pattern on a background of pewter and brass.

With simple grammatical function, both *the coffers* and *brass* are CBs of (s8), which is therefore treated by our script as not having a well-defined CB. As a result, the pronominalization of a non-CB, *premiere partie marquetry*, is not counted as a violation. (s8), however, becomes a violation with GFTHERELIN, since *the coffers* becomes its only CB.

With the IF instantiation, the percentage of violations of Rule 1 (3.7%) is nonsignificantly greater than the percentage with simple grammatical function (3.5%). The percentage of violations with the two IS instantiations (7.4%) is significantly worse (at the .01 level) than with simple grammatical function (6.7% and 7.1%, respectively).

Table 13 shows that using GFTHERELIN has a positive effect on the number of violations of Rule 2 (BFP). Whereas with simple grammatical function, none of the instantiations with indirect realization verifies Rule 2 (BFP) by the Page rank test, with GFTHERELIN the IS instantiation does (although only at the .05 level), as does IF if PRO2s are treated as CF realizations. (All direct realization instantiations still verify the rule.) The main reason for this change is a significant reduction in the percentage of RSH with GFTHERLIN, especially for the IF and IS instantiations: With IF we see a reduction in RSH from 9.7% to 7.6%; with IS, from 14.6% to 11.4%. With the DS instantiation the percentage of RET and RSH is about twice what we find with the DF instantiation, just as we observed with simple grammatical function ranking, but otherwise the results are pretty similar to those with DF. With the IF and the IS instantiations, we get small but significant increases in CON and RET and a reduction in RSH. In Table 14, we report the complete percentages for this instantiation, for comparison with other ranking functions.

The change to GFTHERELIN hardly affects the relative percentages of cheap and expensive transitions, so the results concerning Rule 2 (Strube and Hahn) do not change.

The IS instantiation with GFTHERELIN ranking is the one in which all three claims are verified without need to treat PRO2s as CF realizations, even though Rule 2 is

verified with this instantiation only at the .05 level. We will therefore concentrate on this instantiation when making comparisons with the other ranking variants.

4.4.2 Linear Order. Among the ranking functions alternative to grammatical function, perhaps the simplest is the one that ranks CFs in the order of their occurrence in the utterance, from left to right. This ranking function was explicitly proposed by Rambow (1993) to account for facts about scrambling in German, and effects of order of mention have been observed by, among others, Gernsbacher and Hargreaves (1988), Gordon, Grosz, and Gillion (1993), and Stevenson, Crawley, and Kleinman (1994).

Using linear order instead of GFTHERELIN has no effect at all on Constraint 1, as one would expect, since all that matters for the constraint to be verified is whether discourse entities are mentioned in successive utterances and whether the ranking function is total. However, no significant differences were observed with Rule 1 (GJW 95), either: with IS, we find +378, −30, with linear order, as opposed to +377, −31, with GFTHERELIN.[41] This is because linear order is a very good approximation of grammatical function in English: Subjects tend to occur in first position, objects in second position, etc. The one claim for which the differences are significant is Rule 2 (BFP): with IS, enough CON become RET and enough SSH become RSH that Rule 2 is no longer verified even at the .05 level. (The rule is still verified in the DF and the DS instantiations.)

All in all, these results are not grounds to argue that linear order is a better ranking function than GFTHERELIN;[42] however, because the differences are so small, they also suggest that linear order (which is far easier to compute) might be a good approximation of grammatical-function ranking for practical applications working with English.

4.4.3 Information Structure. Replacing GFTHERELIN with the ranking function proposed by Strube and Hahn (1999) (henceforth, STRUBE-HAHN): rank hearer-old entities more highly than inferrables, and these higher than hearer-old entities—cannot and did not lead to different results for Strong C1, for the reasons already discussed for linear-order ranking. Less expected was the fact that—again, just as in the case of linear order—we didn't find any significant differences with Rule 1 (GJW 95), either, although with the IF and IS instantiations, we find one more violation than with GFTHERELIN.[43] This doesn't mean that the exact same utterances are violations in both cases, rather, that the differences "balance out". We already saw one example in which STRUBE-HAHN ranking results in a violation of Rule 1, whereas GFTHERELIN ranking doesn't: This is the first sentence in (10), illustrating the kind of situations in which a partial ranking may result in two CBs. We repeat that sentence in (21) and include the preceding sentence:

(21)   (s67) An inventory of Count Branicki's possessions made at his death
          describes both the corner cupboard and the objects displayed on its
          shelves: a collection of mounted Chinese porcelain and clocks, some
          embellished with porcelain flowers.

---

41 With IF the difference goes the other way: +463, −18, for GFTHERELIN, +463, −19, for linear order. There are no differences at all with DF and DS.

42 This point is reinforced by a number of results from Gordon and collaborators (e.g., Gordon, Grosz, and Gillion 1993; Gordon et al. 1999) suggesting that hierarchical position in the parse tree is a better predictor of salience than linear order, as well as by results suggesting that for a range of languages, linear order is much less effective–see, for example, Prasad and Strube (2000) for Hindi.

43 We discuss the results only with the version of Rule 1 proposed by Grosz, Joshi, and Weinstein (1995).

Table 15
Transition percentages for IS with STRUBE-HAHN ranking.

	Museum	Pharmaceutical	Total (Percentage)
Establishments	47	60	107 (16.0%)
Continuations	39	55	94 (14.1%)
Retain	50	53	103 (15.4%)
Smooth Shift	18	26	44 (6.6%)
Rough Shift	33	27	60 (9.0%)
ZERO	43	58	101 (15.1%)
NULL	41	119	160 (23.9%)

> (s68) The drawing of the corner cupboard, or more probably an
> engraving of <u>it</u>, must have caught Branicki's attention.

As *the corner cupboard* is in object position, it gets higher ranking in (s67) than *Count Branicki*, which is in NP-modifier position, which—while not explicitly discussed in the centering literature—will presumably fall among the "other" cases. As a result, the cupboard is the CB of (s68), and its pronominalization is predicted by Rule 1. With STRUBE-HAHN ranking, *Count Branicki* is the highest-ranked entity of (s67), and therefore the CB of (s68); hence the violation. Conversely, (22) is an example in which GFTHERELIN and ranking results in a violation of Rule 1, while STRUBE-HAHN ranking doesn't:

(22)   (s88) Christened by his contemporaries as 'the most skillful artisan in
        Paris,' Andrè-Charles Boulle's name is synonymous with the practice of
        veneering furniture with marquetry of tortoiseshell, pewter, and brass.
        (s89) Although he did not invent the technique, Boulle was its greatest
        practitioner and lent his name to its common name: Boulle work.

In this example, *Andrè-Charles Boulle's name*, the subject of (s88), is ranked higher than *Andrè-Charles Boulle*, and is therefore the CB of (s89), in which, however, it is not pronominalized, even though both Boulle and the technique he invented are. Notice that (21) and (22) are almost stereotypical instances of the class of examples that led Sidner (1979) to argue that *two* foci are needed, one for animated entities, and one for the entities acted upon; we return to this issue in Section 5.

The one claim where STRUBE-HAHN ranking makes a clear difference is Rule 2 (BFP). About 20% of RET become CON and about 20% of RSH become SSH. Although we still find more RET than CON and more RSH than SSH, these changes are sufficient to cause Rule 2 (BFP) to be verified at the .01 level in all instantiations considered.[44] The transition percentages with IS and STRUBE-HAHN ranking are presented in Table 15.

Even with IS, however—the instantiation closest to the one proposed by Strube and Hahn—we still find many more expensive transitions (272) than cheap ones (172)

---

44 With DF and DS the number of RET and RSH goes down drastically, so that we do find more CON
   than RET and more SSH than RSH, but we still find more SSH than RET.

and almost three times as many expensive-expensive sequences (137) as cheap-cheap ones (56), so Rule 2 (Strube and Hahn) is not verified.

4.4.4 Summary. Because Strong C1 is the most problematic claim, it was to be expected that the most studied parameter of centering, ranking, would have a smaller impact than the utterance and realization parameters. It is nevertheless interesting that the results for Rule 1 (GJW 95) are virtually identical under the three versions of ranking we considered. More differences can be found for Rule 2 (BFP), which is not verified by any instantiation with linear-order ranking and is verified only by a few instantiations with GFTHERELIN. Adopting STRUBE-HAHN ranking does result in a greater percentage of utterances being classified into one of the "continuous" classes and in a lower probability of Rule 2's (BFP) being falsified. Finally, not even these last changes to parameter settings were sufficient to cause either Rule 1 (Gordon et al.) or Rule 2 (Strube and Hahn) to be verified.

5. Discussion

In this section we first discuss the effects of different parameter settings on the results obtained; we then analyze the claims of centering theory, draw a few theoretical conclusions, and make some suggestions for further work (empirical and theoretical).

5.1 Setting the Parameters
5.1.1 Comparing Instantiations. A central goal of this study was to compare different ways of instantiating centering's parameters, and different versions of its claims, on a single data set, also examining combinations not previously considered (e.g., whether Brennan et al.'s version of Rule 2 would be verified when the parameters were set as suggested by Strube and Hahn and vice versa. Our first interesting result in this sense is that if the parameters are set in the most "mainstream" way (the "vanilla" instantiation) only Rule 1 (GJW 95) and Rule 1 (GJW 83) are clearly verified. The results concerning Constraint 1 are especially negative. As with this instantiation only 35% of utterances are continuous—that is, $CF(U_n) \cap CF(U_{n-1}) \neq \emptyset$ (Kibble 2000; Karamanis 2001)—only the weak version of Constraint 1 is verified. Strong C1, the best-known formulation, and the one that in our view best captures the idea of "entity coherence," clearly doesn't hold. Another interesting observation is that if ranking is required only to be partial, some utterances end up with more than one CB: The percentage of such utterances is only 1% with the vanilla instantiation but can be as high as 6% with some instantiations. This is perhaps obvious, but to our knowledge, it had not been previously discussed.

As for Rule 2, with the vanilla instantiation the version proposed by Brennan et al. is verified by a Page rank test, but arguably, the most striking fact about transitions with this instantiation is the prevalence of NULL transitions (47.9%), Establishments (18.8%), and ZEROs (16.7%). All together, the four types of transitions falling under the remit of Rule 2 account for only 16% of utterances, and if Smooth Shifts and Rough Shifts are counted together, with this instantiation there are more Shifts than Retains. Other classifications and versions of the rule do not correlate much better with the observed frequencies: For example, only 39% of entity-coherent transitions (139 out of 357), and 14% of the total, are cheap in the sense of Strube and Hahn (1999) (i.e., $CP(U_{n-1})$ predicts $CB(U_n)$).

These findings concerning the vanilla instantiation should not, however, lead us to conclude that the theory in general is not verified. Our second major finding is that parameters do matter: that is, it is possible to set the parameters in such a way as to

make all three claims verified in a statistical sense. However, because Strong C1 is the claim with the largest percentage of violations, the parameters whose setting matters the most when one is trying to find an instantiation in which all claims are satisfied are those controlling utterance definition and CF realization. Considering a center as realized in an utterance which contains an associative reference to that center is sufficient for Strong C1 to be verified; identifying utterances with sentences instead of finite clauses also has a strong positive effect. With the resulting instantiations (IF and IS), Strong C1 is verified, as well as the two "basic" versions of Rule 1.

We also found, however, that there is a trade-off between Strong C1, on one side, and Rule 1 and Rule 2, on the other: The changes to the utterance and realization parameters just mentioned, while reducing the violations of Strong C1, increase those of Rule 1 and Rule 2 (see, e.g., Table 13). Identifying utterances with sentences, or (to a lesser extent) allowing indirect realization, results in statistically significant increases in the number of violations of Rule 1—up to a total of 7.4% in the IS instantiation (see Figures 2 and 4)—although Rule 1 (GJW 95) and Rule 1 (GJW 83) are so robust that they are still verified, even in these instantiations.[45] These changes to the utterance and realization parameters have an even greater impact on Rule 2 (BFP), which is only weakly verified with the vanilla instantiation. With the IF and IS instantiations and grammatical-function ranking, we find many more RSH than SSH and many more RET than "pure" CON (i.e., without counting Establishments); indeed, in the IS instantiation with GFTHERELIN ranking, RET are the second most common transition. As a result, Rule 2 (BFP) is verified with IS instantiations only at the .05 level, and with IF instantiations only if second-person pronouns are counted as realizations of CFs. On the positive side, with these instantiations a much greater percentage of utterances (45%) are classified as either CON, RET, SSH, or RSH, and a further 16% as EST.

These results can be further strengthened by making one last change to the parameters: adopting the ranking function proposed by Strube and Hahn (1999) instead of GFTHERELIN. With this instantiation, Rule 2 (BFP) is verified at the .01 level, rather than only at the .05 level. This is because although the STRUBE-HAHN ranking function has no effect on Strong C1 (obviously) or R1 (more surprisingly), it does result in some of the RET becoming CON and some of the SSH becoming RSH. Even though we still find more RET than CON and more RSH than SSH, these changes are enough to cause Rule 2 (BFP) to be verified at the .01 level with the IS instantiation. Strube and Hahn's own version of Rule 2 still isn't verified, but this version of the rule is not verified by any of the instantiations we evaluated. In other words, with the IS or IF instantiation and STRUBE-HAHN ranking, all three claims of the theory are verified at the .01 level.

The final observation concerning parameter settings is that issues not widely discussed in the centering literature had a greater impact on the theory's claims in our experiments than parameters such as the choice of ranking function or the definition of previous utterance. Many of these issues, such as the treatment of second-person pronouns and of empty categories, have to do with the general issue of which entities should be included in the CF list. Considering second-person pronouns to be realizations of discourse entities is enough to make Strong C1 satisfied; we also found that

---

45 Perhaps the most spectacular demonstration of the trade-off between Strong C1 and Rule 1 can be seen with the versions of the theory that adopt the definitions of CB proposed by Gordon, Grosz, and Gillion (1993) and Passonneau (1993). (These instantiations are not discussed in this article, but can be examined on the companion Web site.) By adopting a particularly restrictive definition of CB, these versions succeed in reducing (indeed, eliminating, in the case of Passonneau) the violations of Rule 1, but the price is that only a very few utterances have a CB.

a number of extensions to the definition of utterance, such as the inclusion of relative clauses and nonfinite clauses, led to much worse results unless reduced relative clauses and nonfinite clauses were taken to include traces linking these clauses to the clause in which they were embedded.

5.1.2 Minimizing Violations Should Not Be the Overriding Goal. We said in Section 3 that we don't think that minimizing violations should be the only factor taken into account in deciding how to set parameters. Some violations are best accepted and explained in terms of the interaction of centering preferences with other preferences. (See below.)

Special care is needed when alternative definitions are supported by cross-linguistic evidence or by the results of psychological studies. In the case of ranking, although we didn't find any significant differences between grammatical-function ranking and linear-order ranking for English, one should keep in mind that such differences have been found for other languages, especially more free-order ones. Prasad and Strube (2000), for example, found that in Hindi the difference between grammatical-function and linear-order ranking is significant, and Strube and Hahn (1999) found significant differences between grammatical function and information structure in German. Conversely, before taking the evidence for a slight advantage of STRUBE-HAHN ranking over grammatical-function ranking as conclusive, one needs to supplement our studies with psychological experiments reconciling these results with numerous results indicating the important role played by grammatical function, especially subjecthood (among others, Hudson, Tanenhaus, and Dell [1986]; Gordon, Grosz, and Gillion [1993]; Brennan [1995]). Information structure has also been found not to be appropriate for languages including Greek, Hindi, and Turkish (Turan 1998; Prasad and Strube 2000; Miltsakaki 2002). Similar considerations apply to the definition of previous utterance, since we saw that a considerable amount of psychological evidence supports treating adjuncts as embedded, at least when the syntactically embedded clause is at the end of the sentence (Cooreman and Sanford 1996; Pearson, Stevenson, and Poesio 2000).

In the case of the definition of utterance, our results indicate that identifying utterances with sentences, rather than finite clauses, leads to results much more consistent with the claimed preference for discourses to be entity-coherent. While this result is likely to be useful for a number of reasons and for different types of applications (e.g., text planners), we believe that further empirical and theoretical work is needed before conclusions can be reached about when the local focus is updated. For one thing, most analyses of discourse structure—for example, rhetorical structures theory (Mann and Thompson 1988)–view clauses as the basic unit of discourse in written text. And in spoken dialogue one can hardly find any complete sentences; in this case, the update unit is much more likely to be a prosodic phrase of some sort.

5.2 The Claims of Centering, Revisited
5.2.1 Centering, Pronominalization, and Salience. One clear result of this work is that centering's claims about pronominalization—at least, those expressed by the versions of Rule 1 proposed in Grosz, Joshi, and Weinstein (1995, 1983)–are very robust. Rule 1 (GJW 95) and Rule 1 (GJW 83) are verified with all parameter instantiations, and in a very convincing way: In the instantiations we considered, the percentage of violations of Rule 1 (GJW 95) never exceeds 8% of the total number of utterances.

On the other hand, one should keep in mind that these two versions of Rule 1 make very weak claims about pronominalization. All that Rule 1 (GJW 95) says is that *if* we decide to pronominalize, *then* we should pronominalize the CB. This formulation doesn't address the real problem for a theory of pronominalization or, more generally,

of NP form decision, which is to decide when a discourse entity should be realized as a pronoun (Henschel, Cheng, and Poesio 2000). And our results also indicate that simply strengthening Rule 1 to the form "pronominalize the CB," which can be seen as a generalization of the proposals in Gordon, Grosz, and Gillion (1993), would be a very bad idea: Between 50% (with $u = f$) and 60% (with $u = s$) of mentions of the CB are not realized using a pronoun, and conversely, between 30% and 40% of personal pronouns are not realizations of the CB. Examples like (12) illustrate one type of situation in which a mismatch between the CB and pronominalization may occur: By having been mentioned often in a discourse, a discourse entity may become sufficiently salient (at the global level) to justify pronominalization even when it is not the CB.[46] These observations suggest that the decision to pronominalize does not depend only on whether a discourse entity is the CB but must involve a number of further constraints and preferences.[47]

5.2.2 CT as a Theory of Coherence: Constraint 1. Another result of this work is that the validity of centering's claims about local coherence—Constraint 1 and Rule 2— depends on the choice of the parameters to a much greater extent than is the case for the claims about pronominalization. Strong C1 does not hold for the vanilla instantiation, although it does hold for any instantiation in which the implicit anaphoric component of bridging references is treated as an indirect realization and for many instantiations in which utterances are identified with sentences. But even under the most favorable parameter instantiations, there are many more exceptions to Strong C1 (between 20% and 25% of the total number of utterances) than we find even with the instantiations which are worse for Rule 1 (7–8%). While the weak version of C1, requiring only that there is at most one most salient entity per utterance, does hold even with the vanilla instantiation and does capture the claim that utterances with a unique CB are easier to process, a central aspect of centering since Joshi and Kuhn (1979) and Joshi and Weinstein (1981), it says nothing about entity coherence's being what ensures local coherence.

Further light on entity coherence is shed by recent work on text planning, particularly by Karamanis (2003), which suggests that when all alternative ways of extracting a text plan from the propositions expressed by texts such as those we are studying are considered, the actual ordering found in the texts tends to be in greater agreement with centering's preferences about entity coherence than with most of its alternatives. After extracting the propositions[48] expressed by texts in the museum subdomain of our corpus, Karamanis determined that although the sequence actually found in such texts is not optimal as far as minimizing the violations to entity coherence (with the instantiation he considers, more than 50% of the utterances violate Strong C1), approximately 70% of the alternative orderings introduce even more violations.

If we accept that the texts in our corpus are coherent, these results suggest that there must be other ways of achieving local coherence, apart from what we have been calling here "entity coherence". An obvious candidate for an additional, or

---

46 The role of global focus in the interpretation of pronouns needs further study. A few preliminary observations can be found in Hitzeman and Poesio (1998).
47 The discrepancy between pronominalization and CB-hood in our corpus is analyzed in more detail by Henschel, Cheng, and Poesio (2000), who propose an algorithm for pronominalization that takes into account factors such as the presence of distractors matching the CB's agreement features that may lead to the decision not to pronominalize, as well as factors that may result in the pronominalization of a non-CB. The algorithm achieves an accuracy of 87.8% in the museum domain.
48 More precisely, the lists of CF realized by each utterance with a DF instantiation, representing that utterance's arguments.

alternative, coherence-inducing device are rhetorical relations. Indeed, the claim that "entity" coherence needs to be supplemented by "relational" coherence can be found as far back as Kintsch and van Dijk (1978) and Hobbs (1979). This view is supported by an analysis of our data. With the u=f instantiations, we find in the pharmaceutical subdomain many examples in which successive utterances do not mention the same entities, but the connection between clauses is explicitly indicated by connectives, as in (23):

(23)   (u1) This leaflet is a summary of the important information about Product A.
        (u2) If you have any questions or are not sure about anything to do with your treatment,
        (u3) ask your doctor or your pharmacist.

A more complex case are utterances in the museum subdomain that do not refer to any of the previous CFs because they express generic statements about the class of objects of which the object under discussion is an instance, or utterances that make a generic point that will then be illustrated by a specific object. In (24), (u2) gives background concerning the decoration of a cabinet:

(24)   (u1) On the drawer above the door, gilt-bronze military trophies flank a medallion portrait of Louis XIV. (u2) In the Dutch Wars of 1672–1678, France fought simultaneously against the Dutch, Spanish, and Imperial armies, defeating them all. (u3) This cabinet celebrates the Treaty of Nijmegen, which concluded the war.

While the analysis of such cases in terms of rhetorical relations is more complex, it seems clear to us that an analysis in terms of underlying semantic connections between events or propositions is more perspicuous than one in terms of entity coherence. While it is true that some of these violations could be fixed by adopting a broader notion of bridging reference–for example, in (24) we might treat *France* as a bridge to *Louis XIV*—this wider notion of bridging reference has proven to be very difficult to identify in a reliable way.

Now, given that in an RST-style analysis, every discourse unit is connected by at least one rhetorical link to at least another discourse unit, one might wonder whether "entity coherence" is still needed once "relational coherence" is introduced. However, Knott et al. (2001) convincingly argue that in RST, complete connectivity is usually achieved by introducing relations such as "Elaboration" that, when looked at closely, turn out really to be attempts to capture a notion of entity coherence. This work on rhetorical relations is coming to a position symmetrical to our own: that a purely relational account is not sufficient, and a separate theory of entity coherence is necessary (Knott et al. 2001).[49]

5.2.3 Topic Continuity: Rule 2. Rule 2–stating a preference not just to keep talking about the same objects, but to preserve their relative ranking—also seems much less robust than Rule 1, irrespective of its formulation and of the instantiation.

---

49 The respective role of entity coherence, relational coherence, and other forms of coherence in the examples in our corpus is studied in more detail in Oberlander and Poesio (2002).

As already noted, one of the most interesting observations about this aspect of the theory concerns the classification of utterances used to formalize it (at least in the earlier versions of the theory). With pretty much all parameter instantiations that we tested, two of the most common transitions were the NULL transition (between two utterances neither of which has a CB), previously considered only in Passonneau (1998), and the ZERO transition (from an utterance with a CB to one without), which as far as we can see has never been discussed before. Indeed, with the vanilla instantiation, 84% of all utterances are either NULL, ZERO, or EST and therefore fall outside the scope of Rule 2 in almost all its formulations. The question raised by this finding is whether the theory has to be extended to cover such cases or whether they have to be accounted for by other components of an overall theory of discourse (see below).

Three versions of Rule 2 were tested in some detail.[50] The version of Rule 2 from Grosz, Joshi, and Weinstein (1995), formulated in terms of sequences and stating a preference for sequences of CON over sequences of RET over sequences of SHIFT (which we tested by counting the number of sequence pairs), suffers from the problem that even with the "best" instantiations, fewer than one-third of sequence pairs involve the same transition, and even fewer are sequences of the transitions considered by Grosz et al. Even in the instantiation which yields the best results for Rule 2 (BFP), IS with STRUBE-HAHN ranking, only 13% of sequence pairs are of the form CON-CON/RET-RET/SH-SH, and altogether only 28% of sequence pairs involve only transitions considered by Grosz et al. Keeping in mind that Rule 2 (GJW 95) applies only to a minority of sequence pairs, we do find that with IS and STRUBE-HAHN ranking, the number of CON-CON sequences (37) slightly exceeds the number of RET-RET (35), which in turn exceeds the number of SH-SH (19, of which 16 are RSH-RSH). This doesn't hold with GFTHERELIN ranking, where RET-RET exceeds CON-CON even if we treat EST as a type of CON; we find no significant difference between the IF and the IS setting.

Rule 2 (BFP), formulated in terms of single transitions, accounts for larger percentages of the data (single utterances) and was found to be verified both with the vanilla instantiation and with the "best" instantiations. However, we still observed a large percentage of NULL transitions with most instantiations; we also found more RET than CON and more RSH than SSH in most instantiations in which utterances are identified with sentences or allow for indirect realization.[51]

Finally, Strube and Hahn's preference for sequences of cheap transitions over sequences of expensive ones isn't verified by any of the instantiations we tested; indeed, in all instantiations we studied we found more expensive transitions than cheap ones, meaning that the CP of one utterance generally doesn't predict the CB of the next.

These mixed results are in line with those of psychological experiments, which so far haven't found clear evidence supporting the claim that, say, Continuations are easier to process than Shifts, let alone Retains (Gordon, Grosz, and Gillion 1993)

---

50 As noted earlier, an earlier version of Kibble's proposal was also tested; the results can be viewed on the companion Web site.
51 CON can be made the most frequent transition by merging EST and CON. We found, however, that this merging leads to worse results as far as the correlation between the classification of transitions and two of the linguistic phenomena for which the classification has been used: predicting the form of subject NPs, and predicting segment boundaries. These results are discussed in the technical report.

5.3 Theoretical Consequences

While proposing modifications of centering is beyond the scope of this article, we believe our results do have broad theoretical consequences worthy of further exploration.

5.3.1 Clarification of the Claims and Identification of Further Parameters. Apart from comparing different ways of setting the parameters already discussed in the literature, our work had the more fundamental goal of clarifying the claims of centering theory by identifying aspects that need to be made more precise. Our study raised a number of questions about the definitions of the concepts used in centering not previously examined in the literature or only discussed in passing.

Many of these questions have to do with realization, one of the least studied aspects of the theory. One such question is the status of entities realized as second-person pronouns. Our results indicated that if PRO2s are not considered realizations of CF, or if we treat them as R1-pronouns, we find many more violations of Strong C1 and Rule 1, respectively (although both claims are still verified). We also saw that the results concerning Constraint 1 and Rule 1 depended on whether reduced relative clauses and non-finite VPs were assumed to contain traces and whether or not these traces were assumed to be R1-pronouns. More generally, we identified the need for a clear definition of "R1-pronoun": that is, whether we should include traces in relative clauses, the implicit anaphoric elements of bridging references, and demonstrative pronouns among the "pronouns" to which Rule 1 applies. This question isn't mentioned in the literature we know of, yet our results indicate that, for example, treating the implicit anaphoric elements of bridging references or second-person pronouns as R1-pronouns is a very bad idea.

Some of the issues raised by this study are relevant only for certain parameter instantiations. One example is the specification of grammatical function ranking beyond the simplest cases: for example, whether postcopular NPs in *there*-clauses should be treated as subjects or objects (our results suggest the former) or how nominal modifiers should be ranked (we treated them as adjuncts). An issue for instantiations in which utterances are identified with finite clauses is what the previous utterance is when an embedded finite clause is in the middle of another finite clause, rather than at the end, as in the following example, from the *Guardian*:

(25)   But Hutchinson, who appointed Ranieri last season, today said that he
        spent 30 minutes with the Italian after the Blackburn match and that
        resignation was never an issue.

5.3.2 Separating Entity Coherence from CB Uniqueness. Starting with Brennan, Friedman, and Pollard (1987) and, more recently, Beaver (2004) and Kibble (2001) there have been attempts to "unpack" some of the original preferences proposed by centering. We feel this work has greatly helped our understanding of the theory and believe that it would be similarly useful to unpack Constraint 1 into two separate claims, as well: one about uniqueness of the CB, and one about entity coherence.

The first function of (both versions of) Constraint 1 is to claim that the CB is unique. We will call this claim **CB uniqueness**:

CB Uniqueness: Utterances have at most one CB.

We have argued throughout the article that Strong Constraint 1 has a second function as well: to express a preference for utterances that do not occur at the beginning of a

segment to mention at least one of the objects included in the previous utterance. Following Kibble (2000) and Karamanis (2001), we will call this second half of Constraint 1 **(entity) continuity**:

(Entity) Continuity: $CF(U_{i-1}) \cap CF(U_i) \neq \emptyset$

Weak C1 is CB uniqueness, whereas Strong C1 is CB uniqueness plus continuity.

5.3.3 A Hybrid View of Coherence. One clear conclusion suggested by our results is that entity-based accounts of coherence need to be supplemented by accounts of other factors that induce coherence at the local level. The most direct way to do this would be to include in continuity a longer list of factors that may link an utterance to its previous one and claim that in order for an utterance to be "locally coherent," at least one of these factors must be present. The resulting claim would take a form along the following lines:

Hybrid Continuity For every utterance U_i, at least one of the following must hold:

1. $CF(U_{i-1}) \cap CF(U_i) \neq \emptyset$;
2. There is a rhetorical relation RR such that $RR(U_{i-1},U_i)$.[52]
3. U_{i-1} and U_i are temporally coherent in the sense, for example, of Kameyama, Passonneau, and Poesio (1993).
4. . . . (other factors)

A more sensible approach, especially as we don't yet know all the factors affecting coherence, would be to be more explicit about the scope of centering theory, viewing it not as a comprehensive account of "local coherence," but only as an account of the contribution of entity coherence to local coherence. In other words, we could view (Entity) Continuity as only one among the preferences holding at the discourse level. A natural way to formalize this would be to include Entity Continuity among a set of constraints like those proposed by Beaver, which would also have to include further constraints specifying preferences for rhetorical and temporal coherence.

5.3.4 CB Uniqueness. We saw in Section 4 that it's fairly easy to fix the problem of utterances' violating Weak C1, or CB uniqueness: All that is needed is to strenghten the requirements on the ranking function and require it to be total, which in turn can be easily done by adding a disambiguation factor to ranking functions that aren't total like grammatical function. Before doing this, however, we should ask whether this is the conclusion we should draw from the finding that CB uniqueness will be violated with partial ranking functions—or if instead we should allow for utterances to have more than one CB.

When multi-CB utterances such as (10) are considered, it is not immediately obvious that one discourse entity (*the corner cupboard*) is more salient than the other (*Branicki*), especially since neither occupies a particularly salient position either in the previous utterance (u227) or in the current one (u229). Notice also that both entities have been mentioned before; and furthermore, one of them is animate (*Branicki*), the

---

52 This formulation was intentionally designed in such a way as to finesse the issue of whether RR should be an informational-level relation between the eventualities expressed by the utterances or a genuine rhetorical relation between the speech acts performed by them.

other inanimate (*the cupboard*). In these respects, these examples are reminiscent of the examples that led Sidner (1979) to argue for two foci—sentences with one animate entity (typically in agent position) and an inanimate one (typically in theme position), like *Mortimer sold the book for 10 cents* or *Mary took a nickel from her toy bank yesterday*. Although the results from studies such as Gordon, Grosz, and Gillion (1993) suggest that when two animate entities are considered, only one tends to show RNP effects, we are not aware of any experiment testing materials like those discussed by Sidner.

The hypothesis that topicality is not restricted to one entity per utterance has been advanced by a number of researchers, although it is perhaps most clearly associated with the work of Givon (1983). Within the centering literature, abandoning the claim that we call "CB uniqueness" has been suggested by Gundel (1998) and, more radically, in work such as Strube (1998), Gordon and Hendrick (1999), and Tetreault (2001) in which the whole notion of CB is abandoned.

As seen in Section 2, the primary motivation for CB uniqueness is complexity-theoretic arguments: Inference in monadic logics is less expensive than in normal logics (Joshi and Kuhn 1979; Joshi and Weinstein 1981). Grosz and colleagues's linguistic evidence for CB uniqueness is contrasts like those in (3), showing that failing to pronominalize certain entities (*Susan*, in that example) is a more serious problem than failing to pronominalize others (*Betsy*). This claim is further supported by the evidence concerning the Repeated Name Penalty (Gordon, Grosz, and Gillion 1993). However, the RNP is observed only in a subset of the cases that would be considered as CB mentions according to the definition provided by Constraint 3, and in the example we are discussing (10), neither *Branicki* nor *the cupboard* occur in (u229) in a position that would be subject to RNP effects according to Gordon et al. In other words, (some) evidence used by Grosz et al. in support of CB uniqueness cannot be used to argue that (u229) in (10) has a single CB. This evidence is also consistent with a different solution to the problem raised by examples like (10): Instead of attempting to preserve CB uniqueness by requiring the ranking function to be total, one could abandon CB uniqueness, as suggested in Givon (1983) and Gundel (1998). In both cases, we would need a separate theoretical account of RNP effects. More empirical evidence is needed on this issue.[53]

5.3.5 Variety. The third conclusion suggested by our results is that ensuring variety seems to be as important a principle in discourse production as maintaining coherence. This is suggested, first of all, by the fact that only slightly over half of CBs are realized as R1-pronouns. It is also the case that CBs are hardly ever continued for more than two or three utterances; that the same discourse entity is very unlikely to be realized using the same type of NP twice in a row (even with pronouns, we only have 58 pronoun-pronoun sequences—26% of the total); and that two-thirds of all transition sequences involve two different transitions. In fact, we wonder whether that the repeated name penalty observed by Gordon et al. might not be an instance of this more general phenomenon.

---

53 One way to reconcile the different findings would be to use different conceptual tools to characterize the connection between subsequent utterances. Each utterance satisfying Continuity would have one or more links to the previous utterance, that we might call **centers of coherence**; Entity Continuity would then become a preference for the set of centers of coherence to be nonempty. In particular situations, may be experimentally identified using the RNP as a test, one of the centers of coherence may acquire a particular status, leading to a preference for pronominalization. We may call this center the **center of salience**, say. It would also be interesting to examine the connection between a solution along these lines and Sidner's solution involving two foci.

References

Alshawi, Hiyan. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge.

Arnold, Jennifer E. 1998. *Reference Form and Discourse Patterns*. Ph.D. thesis, Stanford University, Stanford CA.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project, Montreal, Canada. In *Proceedings of the 36th ACL*, pages 86–90.

Beaver, David. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy*, 27(1):3–56.

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

Brennan, Susan E., Marilyn W. Friedman, and Charles J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th ACL*, Stanford, CA, pages 155–162, June.

Byron, Donna and Amanda Stent. 1998. A preliminary model of centering in dialog. In *Proceedings of the 36th ACL*, Montreal, Canada.

Caramazza, Alfonso, Ellen Grober, Catherine Garvey, and Jack Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior*, 16:601–609.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, and topics. In C. Li, editor, *Subject and Topic*. Academic Press, New York, pages 25–76.

Chinchor, Nancy A. and Beth Sundheim. 1995. Message understanding conference (MUC) tests of discourse processing. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford, CA.

Clark, Herbert H. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420, Cambridge University Press, London and New York.

Cooreman, Ann and Tony Sanford. 1996. Focus and syntactic subordination in discourse. Research Paper no. RP-79, University of Edinburgh, HCRC.

Cote, Sharon. 1998. Ranking forward-looking centers. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*. Oxford, University Press, Oxford, chapter 4, pages 55–70.

Dale, Robert. 1992. *Generating Referring Expressions*. MIT Press, Cambridge, MA.

Di Eugenio, Barbara. 1998. Centering in Italian. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*. Oxford University Press, Oxford, chapter 7, pages 115–138.

Di Eugenio, Barbara, Johanna D. Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th ACL*, Madrid.

Fox, Barbara A. 1987. *Discourse Structure and Anaphora*. Cambridge University Press, Cambridge.

Gernsbacher, Morton A. and David Hargreaves. 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27:699–717.

Giouli, Paraskevi. 1996. Topic chaining and discourse structure in task-oriented dialogues. Master's thesis, Linguistics Department, University of Edinburgh.

Givon, Talmy, editor. 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Benjamins, Amsterdam and Philadelphia.

Gordon, Peter C., Barbara J. Grosz, and
Laura A. Gillion. 1993. Pronouns, names,
and the centering of attention in
discourse. *Cognitive Science*, 17:311–348.

Gordon, Peter C. and Randall Hendrick.
1999. The representation and processing
of coreference in discourse. *Cognitive
Science*, 22:389–424.

Gordon, Peter C., Randall Hendrick,
Kerry Ledoux, and Chin L. Yang. 1999.
Processing of reference and the structure
of language: An analysis of complex noun
phrases. *Language and Cognitive Processes*,
14(4):353–379.

Grosz, Barbara J. 1977. *The Representation and
Use of Focus in Dialogue Understanding*.
Ph.D. thesis, Stanford University,
Stanford, CA.

Grosz, Barbara J., Aravind K. Joshi, and
Scott Weinstein. 1983. Providing a unified
account of definite noun phrases in
discourse. In *Proceedings of ACL-83*,
Cambridge, MA, pages 44–50.

Grosz, Barbara J., Aravind K. Joshi, and
Scott Weinstein. 1986. Towards a
computational theory of discourse
interpretation. Unpublished manuscript.

Grosz, Barbara J., Aravind K. Joshi, and
Scott Weinstein. 1995. Centering: A
framework for modeling the local
coherence of discourse. *Computational
Linguistics*, 21(2):202–225.

Grosz, Barbara J. and Candace L. Sidner.
1986. Attention, intention, and the
structure of discourse. *Computational
Linguistics*, 12(3):175–204.

Gundel, Jeanette K. 1998. Centering theory
and the givenness hierarchy: Towards a
synthesis. In M. A. Walker, A. K. Joshi,
and E. F. Prince, editors, *Centering Theory
in Discourse*. Oxford University Press,
Oxford, chapter 10, pages 183–198.

Gundel, Jeanette K., Nancy Hedberg, and
Ron Zacharski. 1993. Cognitive status and
the form of referring expressions in
discourse. *Language*, 69(2):274–307.

Hawkins, John A. 1978. *Definiteness and
Indefiniteness*. Croom Helm, London.

Heim, Irene. 1982. *The Semantics of Definite
and Indefinite Noun Phrases*. Ph.D. thesis,
University of Massachusetts at Amherst.

Henschel, Renate, Hua Cheng, and
Massimo Poesio. 2000. Pronominalization
revisited. In *Proceedings of 18th COLING*,
Saarbruecken, August.

Hitzeman, Janet, Alan Black, Paul Taylor,
Chris Mellish, and Jon Oberlander. 1998.
On the use of automatically generated
discourse-level information in a
concept-to-speech synthesis system. In
*Proceedings of the International Conference on
Spoken Language Processing (ICSLP98)*,
Sydney, Australia.

Hitzeman, Janet and Massimo Poesio. 1998.
Long-distance pronominalisation and
global focus. In *Proceedings of
ACL/COLING*, vol. 1, pages 550–556,
Montreal.

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

Hudson, Susan B., Michael K. Tanenhaus,
and Gary S. Dell. 1986. The effect of the
discourse center on the local coherence of
a discourse. In *Proceedings of the Eight
Annual Meeting of the Cognitive Science
Society*, Amherst, MA, pages 96–101.

Hudson-D'Zmura, Susan and Michael K.
Tanenhaus. 1998. Assigning antecedents
to ambiguous pronouns: The role of the
center of attention as the default
assignment. In M. A. Walker, A. K. Joshi,
and E. F. Prince, editors, *Centering Theory
in Discourse*. Oxford University Press,
Oxford, pages 199–226.

Hurewitz, Felicia. 1998. A quantitative look
at discourse coherence. In M. A. Walker,
A. K. Joshi, and E. F. Prince, editors,
Centering Theory in Discourse. Oxford
University Press, Oxford, pages 273–291.

Joshi, Aravind K. and Steve Kuhn. 1979.
Centered logic: The role of entity centered
sentence representation in natural
language inferencing. In *Proceedings of the
IJCAI*, Tokyo, pages 435–439.

Joshi, Aravind K. and Scott Weinstein. 1981.
Control of inference: Role of some aspects
of discourse structure–centering. In
Proceedings of the IJCAI, Vancouver, CA,
pages 385–387.

Kameyama, Megumi. 1985. *Zero Anaphora:
The Case of Japanese*. Ph.D. thesis, Stanford
University, Stanford, CA.

Kameyama, Megumi. 1986. A
property-sharing constraint in centering.
In *Proceedings of the ACL-86*, New York,
pages 200–206.

Kameyama, Megumi. 1998. Intra-sentential
centering: A case study. In M. A. Walker,
A. K. Joshi, and E. F. Prince, editors,
Centering Theory in Discourse. Oxford
University Press, Oxford, chapter 6,
pages 89–112.

Kameyama, Megumi, Rebecca Passonneau,
and Massimo Poesio. 1993. Temporal
centering. In *Proceedings of the 31st ACL*,
pages 70–77, Columbus, OH.

Kamp, Hans and Uwe Reyle. 1993. *From
Discourse to Logic*. Reidel, Dordrecht.

Karamanis, Nikiforos. 2001. Exploring
entity-based coherence. In *Proceedings of
the Fourth CLUK*. University of Sheffield,
pages 18–26.

Karamanis, Nikiforos. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, Informatics, University of Edinburgh.

Karttunen, Lauri. 1976. Discourse referents. In J. McCawley, editor, *Syntax and Semantics 7—Notes from the Linguistic Underground*. Academic Press, New York, pages 363–385.

Kibble, Rodger. 2000. A reformulation of rule 2 of centering theory. Technical report, ITRI, University of Brighton.

Kibble, Rodger. 2001. A reformulation of Rule 2 of centering theory. *Computational Linguistics*, 27(4):579–587.

Kibble, Rodger and Richard Power. 2000. An integrated framework for text planning and pronominalization. In *Proceedings of INLG*, Mitzpe Ramon, Israel, June.

Kintsch, Walter and Teun van Dijk. 1978. Towards a model of discourse comprehension and production. *Psychological Review*, 85:363–394.

Knott, Alistair, Jon Oberlander, Mick O'Donnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*. John Benjamins, Amsterdam and Philadelphia, pages 181-196.

Lappin, Shalom and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Lascarides, Alex and Nick Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, Daniel. 1999. Instructions for manually annotating the discourse structures of texts. Unpublished manuscript, USC/ISI.

McKeown, Kathy R. 1985. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41.

Miltsakaki, Eleni. 1999. Locating topics in text processing. In *Proceedings of CLIN*, Utrecht, pages 127–138.

Miltsakaki, Eleni. 2002. Towards an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, 28(3):319–355.

Moser, Megan and Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Oberlander, Jon, Mick O'Donnell, Alistair Knott, and Chris Mellish. 1998. Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.

Oberlander, Jon and Massimo Poesio. 2002. Entity coherence and relational coherence: a corpus-based investigation. Paper presented at the Berlin workshop on Topics in Discourse, September.

Passonneau, Rebecca J. 1993. Getting and keeping the center of attention. In M. Bates and R. M. Weischedel, editors, *Challenges in Natural Language Processing*. Cambridge University Press, Cambridge, chapter 7, pages 179–227.

Passonneau, Rebecca J. 1997. Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.

Passonneau, Rebecca J. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*. Oxford University Press, Oxford, chapter 17, pages 327–358.

Passonneau, Rebecca J. and Diane Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*, Columbus, OH, pages 148–155.

Pearson, Jamie, Rosemary Stevenson, and Massimo Poesio. 2000. Pronoun resolution in complex sentences. In *Proceedings of AMLAP*, Leiden.

Pearson, Jamie, Rosemary Stevenson, and Massimo Poesio. 2001. The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In M. Poesio, editor, *Proceedings of the First SEMPRO*, University of Edinburgh.

Poesio, Massimo. 2000. Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of the 2nd LREC*, pages 211–218, Athens, May.

Poesio, Massimo. 2003. Associative descriptions and salience. In *Proceedings of the EACL Workshop on Computational Treatments of Anaphora*, Budapest.

Poesio, Massimo. 2004. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*, Boston, May.

Poesio, Massimo, Florence Bruneseaux, and
    Laurent Romary. 1999. The MATE
    meta-scheme for coreference in dialogues
    in multiple languages. In M. Walker,
    editor, *Proceedings of the ACL Workshop on
    Standards and Tools for Discourse Tagging*,
    pages 65–74.

Poesio, Massimo and Barbara Di Eugenio.
    2001. Discourse structure and anaphoric
    accessibility. In Ivana Kruijff-Korbayová
    and Mark Steedman, editors, *Proceedings
    of the ESSLLI 2001 Workshop on Information
    Structure, Discourse Structure and Discourse
    Semantics*, Helsinki.

Poesio, Massimo and
    Natalia Nygren-Modjeska. 2003. The
    THIS-NP hypothesis: A corpus-based
    investigation. In *Proceedings of DAARC*,
    Lisbon.

Poesio, Massimo and Rosemary Stevenson.
    Forthcoming. *Salience: Theoretical Models and
    Empirical Evidence*. Cambridge University
    Press, Cambridge and New York.

Poesio, Massimo and Renata Vieira. 1998. A
    corpus-based investigation of definite
    description use. *Computational Linguistics*,
    24(2):183–216.

Prasad, Rashmi and Michael Strube. 2000.
    Discourse salience and pronoun
    resolution in Hindi. In *Penn Working
    Papers in Linguistics*, volume 6, University
    of Pennsylvania, Philadelphia,
    pages 189–208.

Prince, Ellen F. 1981. Toward a taxonomy of
    given-new information. In P. Cole, editor,
    Radical Pragmatics. Academic Press, New
    York, pages 223–256.

Prince, Ellen F. 1992. The ZPG letter:
    Subjects, definiteness, and information
    status. In S. Thompson and W. Mann,
    editors, *Discourse Description: Diverse
    Analyses of a Fund-Raising Text*. Benjamins,
    Amsterdam and Philadelphia,
    pages 295–325.

Quirk, Randolph and Sidney Greenbaum.
    1973. *A University Grammar of English*.
    Longman, Harlow, England.

Rambow, Owen. 1993. Pragmatics aspects
    of scrambling and topicalization in
    German. In *Proceedings of the Workshop
    on Centering Theory in Naturally-
    Occurring Discourse*. Institute for
    Research in Cognitive Science,
    Philadelphia.

Sanford, Anthony J. and Simon C. Garrod.
    1981. *Understanding Written Language*.
    Wiley, Chichester, England.

Scott, Donia, Richard Power, and
    Roger Evans. 1998. Generation as a
    solution to its own problem. In
    *Proceedings of the ninth International

Workshop on Natural Language Generation*,
    Niagara-on-the-Lake, CA.

Sgall, Petr. 1967. Functional sentence
    perspective in a generative description.
    Prague Studies in Mathematical Linguistics,
    2:203–225.

Sidner, Candace L. 1979. *Towards a
    Computational Theory of Definite Anaphora
    Comprehension in English Discourse*. Ph.D.
    thesis, Massachusetts Institute of
    Technology, Cambridge.

Siegel, Sidney and N. John Castellan. 1988.
    *Nonparametric Statistics for the Behavioral
    Sciences,* 2nd edition McGraw-Hill, Boston.

Stevenson, Rosemary J., Rosalind A.
    Crawley, and David Kleinman. 1994.
    Thematic roles, focus, and the
    representation of events. *Language and
    Cognitive Processes*, 9:519–548.

Stevenson, Rosemary, Alistair Knott,
    Jon Oberlander, and Scott McDonald.
    2000. Interpreting pronouns and
    connectives: Interactions between
    focusing, thematic roles and coherence
    relations. *Language and Cognitive Processes*,
    15, pages 225–262.

Strube, Michael. 1998. Never look back: An
    alternative to centering. In *Proceedings of
    COLING-ACL*, pages 1251–1257, Montreal.

Strube, Michael and Udo Hahn. 1999.
    Functional centering–Grounding
    referential coherence in information
    structure. *Computational Linguistics*,
    25(3):309–344.

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, 20(2):301–317.

Tetreault, Joel R. 2001. A corpus-based
    evaluation of centering and pronoun
    resolution. *Computational Linguistics*,
    27(4):507–520.

Turan, Umit. 1998. Ranking forward-looking
    centers in Turkish: Universal and
    language-specific properties. In M. A.
    Walker, A. K. Joshi, and E. F. Prince,
    editors, *Centering Theory in Discourse*.
    Oxford University Press, Oxford,
    chapter 8, pages 139–160.

Vallduvi, Enric. 1990. *The Informational
    Component*. Ph.D. thesis, University of
    Pennsylvania, Philadelphia.

van Deemter, Kees and Rodger Kibble. 2000.
    On coreferring: Coreference in MUC and
    related annotation schemes. *Computational
    Linguistics*, 26(4):629–637.

Walker, Marilyn A. 1989. Evaluating
    discourse processing algorithms. In
    Proceedings ACL-89, pages 251–261,
    Vancouver, British Columbia, Canad, June.

Walker, Marilyn A. 1993. Initial contexts and shifting centers. In *Proceedings of the Workshop on Centering*, University of Pennsylvania, Philadelphia.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince. 1998a. Centering in naturally occurring discourse: An overview. In M. A. Walker, A. K. Joshi, and E. F. Prince, editors, *Centering Theory in Discourse*. Oxford University Press, Oxford, chapter 1, pages 1–28.

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince, editors. 1998b. *Centering Theory in Discourse*. Oxford University Press, Oxford.

Webber, Bonnie L. 1978. A formal approach to discourse anaphora. Report no. 3761, BBN, Cambridge, MA.