Centering: a parametric theory and its instantiations

Massimo Poesio
University of Essex
Rosemary Stevenson
University of Durham
Barbara Di Eugenio
The University of Illinois at Chicago
Janet Hitzeman
The 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 of these different ways of setting the parameters on the theory's claims. Doing this required, first of all, to clarify 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: e.g., 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* 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 only holds 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 tradeoff 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 viceversa. Our results suggest that 'entity' coherence–continuous reference to the same entities–must be supplemented at least with an account of relational coherence.

1 MOTIVATIONS

Centering Theory (Joshi and Weinstein, 1981; Grosz, Joshi, and Weinstein, 1983; Grosz, Joshi, and Weinstein, 1995; Walker, Joshi, and Prince, 1998b) is the component of Grosz and Sidner's overall theory of attention and coherence in discourse (Grosz, 1977; Sidner, 1979; Grosz and Sidner, 1986) concerned with *local* coherence and salience, i.e., coherence and salience within a discourse segment.

A fundamental characteristic of Centering is that it is best 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: e.g.,

different definitions of CB can be found in (Grosz, Joshi, and Weinstein, 1983; Grosz, Joshi, and Weinstein, 1995; 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 paper was to explore the extent to which the preferences proposed in Centering are affected by these different ways of instantiating the parameters of the theory. 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 are behavioral experiments, and many aspects of the theory have been in fact 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 only be done by computational means. A corpus-based evaluation has other advantages, as well– among which, 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 reliable 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 paper. (The web site also contains the annotation manual, and a Technical Report expanding this paper 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 paper is organized as follows. We first review the basic concepts of the theory, discussing the three claims on which we focus–Constraint1, Rule 1, and Rule 2–and the parameters they contain. We then discuss how the corpus was annotated, and how the annotation was used to compute violations of the three main claims. In Section §4 we discuss our main results. Our results are discussed in Section §5.

2 CENTERING THEORY AND ITS PARAMETERS

It is not possible to discuss in this paper 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 paper. For more details, we refer the reader to classic references such as (Grosz, Joshi, and Weinstein, 1995; Walker, Joshi, and Prince, 1998b) or the discussion of Centering in (Poesio and Stevenson, To appear).

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*: i.e., 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 already formulated by Chafe (1976) and is backed by empirical evidence such as (Kintsch and van Dijk, 1978; Givon, 1983). In Centering this hypothesis is further strenghtened by proposing that every utterance has a unique 'main link' with the previous utterance: the 'Backward-Looking Center', or 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*: i.e., 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).

These claims about coherence and salience are linked by two further hypotheses:

1 Entity-based theories of coherence are so-called by contrast with RELATION-CENTERED theories of coherence, such as those developed in (Hobbs, 1979; Mann and Thompson, 1988) and used in (Fox, 1987; 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 the Discussion.)

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 literature (Gundel, Hedberg, and Zacharski, 1993) 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) only differ in their (d) sentence, 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 (c)-(f) of the discourse initiated by utterances (3a-b).

(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 (b), all variants should be equally acceptable.

2.2 Terminology and Definitions

*Local Focus, Forward-Looking Centers (*CFs*) 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 'potential discourse foci' (Sidner, 1979) 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 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 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.

Ranking, CP *and* CB We already mentioned two important claim 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', or CP. Ranking is also used to characterize one of the CFs as the BACKWARD-LOOKING CENTER (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 only characterized in intuitive terms, most subsequent work has been based on the definition below (Grosz, Joshi, and Weinstein, 1995), called '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 depend exclusively on ranking and 'previous utterance,' making these parameters crucially important for the framework. [4]

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 between three types of transitions, depending on whether the backward looking center of U_{i-1} is maintained

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.

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.

2.3 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), p. 210). They 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, and by which they are now best known:

Constr. 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 (GJW 95): (Sequences of) continuations are preferred over (sequences of) retains, which are preferred over (sequences of) shifts.

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, 1998; Miltsakaki, 1999; Beaver, 2004), Constraint 1 expresses, first and foremost, the original claim from (Joshi

and Kuhn, 1979; 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; 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, p.3)), preserving 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 1st have *at most one* CB.

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 propose a more restrictive definition of CB (briefly, that the CB is the entity subject to the

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. 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.

RNP–for discussion, see the longer version of the paper) 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 only properly evaluated using that definition. (The results with this instantiation are discussed in the longer version of the paper.)

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 preserving 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 the empirical work just mentioned. 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) found 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 pre-

ferred to the RET transition, which is preferred to the SMOOTH-SHIFT transition (SSH), which is preferred to the ROUGH-SHIFT transition (RSH).

This formulation of Rule 2 depends on a further distinction between two types of SHIFT: SMOOTH SHIFT, when CB(U_n) = CP(U_n) and ROUGH-SHIFT, 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, for utterances 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 only determined 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 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 dis-

tinction between CHEAP and EXPENSIVE transitions ((Strube and Hahn, 1999), p.332):

- A transition pair is CHEAP if the CB of the current utterance is correctly predicted by the CP of the previous utterance, i.e., 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]

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

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–i.e., 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 paper, and the results can be seen in the longer Technical Report accompanying this paper, 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 paper and (Poesio and Stevenson, To appear).

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 is updated after every tensed clause, not after every sentence; and classified tensed clauses into (i) utterances that constitute a 'permanent' update of the local focus, such as coordinated clauses and adjuncts, and (ii) 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 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 semi-controlled 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 proposes 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 only updated after every sentence, and that only the CFs in the main clause are considered when establishing the CB.

Realization Grosz, Joshi, and Weinstein (1995) simply say that the definition of 'U realizes 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 CF in the sense of Hawkins (1978),[8] i.e., 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 towards *the house*. (u2) <u>The door</u> was open.

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; 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) only has a CB (the house) if *the house* is considered to be realized in (u2) by virtue of it being associated with *the door*. To our knowledge, the effect of these alternative notions of realization on the predictions of the theory have not been previously studied, even though theories of focusing such as Sidner's (1979) do allow the (discourse) fo-

8 Associative references are one type of BRIDGING REFERENCE (Clark, 1977).

cus 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 VPs as in (6c):

(6) a. John wanted (to ∅ buy a house).

b. John bought a house (∅ abandoned by its previous occupiers).

c. John bought a house and (∅ promptly demolished it).

If, e.g., the coordinated VP in (6c) is considered a separate utterance, whether or not it contains a realization of *John* is going to determine whether it has a CB or not. To our knowledge, morphologically null elements have only been considered in the Centering literature 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 only contains entities realized as third person 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.

9 According to Walker, Joshi, and Prince (1998a), in the original version of (Grosz, Joshi, and Weinstein, 1995), that 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 realizes as first and second pronouns–would be considered as realizes by an utterance.

(u3) Whilst you are receiving PRODUCT-Z ....

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; Walker, Joshi, and Prince, 1998a) claim that the factors affecting ranking may not be the same in all languages. Nevertheless, most versions of the theory developed since (Kameyama, 1985; Kameyama, 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, and these are ranked more highly than other grammatical positions: SUBJ $\prec$ OBJ $\prec$ OTHERS (see also (Kameyama, 1986; Hudson, Tanenhaus, and Dell, 1986)). Slightly different ranking functions based on grammatical function where proposed by Brennan, Friedman, and Pollard (1987) (who made a further distinction between objects and indirect objects), by Walker, Iida, and 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; e.g., the results of probe experiments by Gernsbacher and Hargreaves (1988) 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 (RNP) for CFs in subject position co-referring 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 non-subject 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; and in turn, these rank higher than HEARER-NEW entities: HEARER-OLD $\prec$ MEDIATED $\prec$ 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 claim (in her dissertation) 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., 1977) as well as work by (Stevenson, Crawley, and Kleinman, 1994; Pearson, Stevenson, and Poesio, 2001a). 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, the type of connective, and animacy (Stevenson, Crawley, and Kleinman, 1994; Stevenson et al., 2000; Pearson, Stevenson, and Poesio, 2001a).

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.

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 - i.e., if the choice is between using a third person singular pronoun to realize a CB or another CF, the CB should be chosen. We also saw that in languages such as Italian, Japanese, and Turkish, the preferred realization of CBs are morphologically null elements (Kameyama, 1986; Walker, Iida, and Cote, 1994; Turan, 1998; Di Eugenio, 1998). But should an utterance of English count as verifying the rule if 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-PRONOUNs, 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, 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 for both 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 found increasingly useful in text planning (McKeown, 1985; Kibble and Power, 2000; Knott et al., 2001; Karamanis, 2003).

We already saw that some predictions of the theory have been tested with psychological techniques. In many of these experiments, differences in processing pronominal references 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 non-pronominal references or references to non-subjects (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-mention 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, none of the previous studies has attempted to analyze in a systematic ways 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 used for this comparison was 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 doing the comparison, the data we used, our annota-

tion 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, i.e., 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 these preferences to interact with other constraints (a point not emphasized enough in the Centering literature). And secondly, 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: i.e., 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. This is how our evaluation was carried out; the tests we used are the Sign test for Constraint

11 In this version of the paper, we assume the CB is defined by Constraint 3. For the results with alternative definitions of CB, see the extended Technical Report or the companion website.

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 his 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 on this issue in the Discussion.

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 these 'violations' actually result in processing difficulties: this can only be determined 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 do the comparison 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 the Discussion.

3.2 The Data

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

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,

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 previously looked at (Oberlander et al., 1998). The SOLE project extended ILEX with concept-to-speech abilities, using linguistic information to control intonation (Hitzeman et al., 1998).

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

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; Passonneau, 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/his corpus according to her/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 about segmentation and about ranking, as discussed below. The annotation followed a detailed manual, available from 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 paper and the web site.

A systematic comparison between 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 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, i.e., 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.

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 ending with a full stop, 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 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, appositive, a parenthetical, etc.
- verbed: whether the unit contains a verb or not.
- finite: for verbed units, whether the verb is finite or not.

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

Marking up sentences proved up to be quite easy; marking up units required extensive annotator training. The agreement on identifying the boundaries of units, using the κ statistic discussed in (Carletta, 1996), was $\kappa = .9$ (for two annotators and 500 units); the agreement on features (2 annotators and at least 200 units) was as follows: utype: κ=.76; verbed: κ=.9; finite: κ=.81. In total, the texts used for the main study contain 505 sentences and more than 1,000 units, including 900 finite clauses.

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 total, the corpus used for this study contains 3345 NPs. These include 586 pronouns, among which 217 third-person personal and possessive pronouns, 23 demonstratives, and 308 second person pronouns; 1290 definite NPs including 554 *the*-nps, 250 possessive NPs, and 391 proper nouns; 1119 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 NPs specifying their syntactic, semantic and discourse properties (Poesio, 2000). Those relevant to the study discussed here include:

- The NP type, cat, with values such as a-np, that-np, the-np, pers-pro, etc. .
- 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 post-verbal objects in copular sentences, such as *This is (a production watch)*), there-obj (for post-verbal 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).

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 scheme). 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, 1998; 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.

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 CFs enter into, including both identity relations and, in order to compute indirect realization, associative relations. 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 these proposals 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 finer-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 allow 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>
 </unit>,
...
</unit>
<ante current="ne549" rel="ident"> <anchor ID="ne547"> </ante>
```

Work such as (Sidner, 1979; Strube and Hahn, 1999), as well as our own preliminary analysis, suggested that indirect realization can play a crucial role in maintaining the CB. However, previous attempts at marking anaphoric information, particularly in the

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

context of the MUC initiative, suggested that while it's fairly easy to achieve agreement on identity relations, marking up bridging references is quite hard; this was confirmed by studies such as (Poesio and Vieira, 1998). For these reasons, and to reduce the annotators' work, we only marked a few types of relations, and we specified priorities. Besides identity (`IDENT`) we only marked up three associative relations (Hawkins, 1978). These relations include set membership (`ELEMENT`), subset (`SUBSET`), and 'generalized possession' (`POSS`), which includes part-of relations as well as ownership relations. We only marked 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.) Furthermore, we specified preferences: 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 looking at 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 limit 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 are 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).

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

Segmentation According to Grosz and Sidner (1986), Centering is only meant to capture preferences within discourse segments. A proper evaluation of the claims of the theory would require therefore a corpus in which discourse segments have been identified. Unfortunately, discourse segments are difficult to identify reliably (Passonneau and Litman, 1993), and Grosz and Sidner (1986) do not provide a specification of discourse intentions explicit enough that can be used to identify the intentional structure of texts– which, according to Grosz and Sidner, determines their segmentation. 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 contains a pronoun in subject position, or a pronoun whose agreement features are not matched by any other CF in the same sentence. We only tested heuristic methods as well, using the layout structure of the texts as a rough indicator of discourse structure. In this paper we only discuss the results with the heuristic proposed by Walker. In the extended Technical Report 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, and 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 CB) according to the particular parameter instantiation chosen, and 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 paper 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-style.

realization: only allow direct realization, or 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 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 who 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 extended report.) The algorithm used to compute the statistics concerning the violations of the claims is fairly straightforward, and we will therefore omit it here; the interested reader can find a discussion in the extended

report. The one additional complication that we need to mention here are 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 only realized by a relative pronoun. What this means in practice is that the 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, realization, and segmentation. After establishing the 'best' values for these parameters, we consider the effect of alternative ranking functions. Additional results are discussed in the extended report. Readers who want to try out instantiations not discussed here should try the companion web site.

4.1 The Vanilla Instantiation

What we call '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 gram-

matical function for ranking, as proposed there and in (Brennan, Friedman, and 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 only counts as 'realized' in an utterance if it is explicitly mentioned. For the purposes of Rule 1, we mainly studied a 'strict' definition of R1-pronoun allowing only personal (and possessive) pronouns and relative pronouns and traces (see the introduction to (Walker, Joshi, and Prince, 1998b), p. 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 embedding NP, 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.

	MUSEUM	PHARMA	TOTAL
Number of utterances:	430	577	1007
(Of which are segment boundaries) :	91	134	225
Number of CFs:	1731	1308	3039

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

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 is likely to be verified with the 'Vanilla' instantiation. Even without counting segment boundaries, Weak C1 is verified by 833 utterances (82.7%) and violated by only 11 (1%): the chance that Weak C1 will not hold with a different sample is $p \leq 0.001$ by the sign test. (We will

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.

	MUSEUM	PHARMA	TOTAL (PERC)
Number of times at least one CF(Un) is realized in Un+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%)
Utterances without a CB but are segment boundary :	67	96	163 (16.2%)
Utterances without a CB :	168	319	487 (48.4%)

Table 2
Utterances and CBs with the Vanilla instantiation.

henceforth write +833, -11 to indicate 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 out of 1007 (34.4%) have exactly one CB, whereas 498 utterances have zero or more than one CB (49.4%). 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 Vanilla 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 a text is coherent: only 35.4% of utterances in our corpus would be 'entity-coherent' i.e., 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

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.

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

The fact that 11 utterances (1%) have *more than one* CB - i.e., they violate Weak C1 as well - is also worth noticing. The reason for this is that in 'classic' Centering ranking is only required to be a partial order (see, e.g., the intro to (Walker, Joshi, and Prince, 1998b), p. 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 ut-

21 It's not clear to us why ranking is only required to be partial, yet the CB is clearly claimed to be unique.

terance 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), 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 the Discussion.

Rule 2 The statistics relevant for Brennan *et al.*'s version of Rule 2 are shown in Table 3.

	MUSEUM	PHARMA	TOTAL
Establishment :	96	95	189 (18.8%)
Continuation :	37	33	70 (6.9%)
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	1007

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

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 only in a limited way. By far the most frequent transition (47.9% of the total) is NULL: follow up an utterance without a CB with a second

22 Neither u228 nor u230 are treated as utterances as they are not finite.

one also without a CB. (Examples include u3, u4, and u5 in (9).) We only found this transition discussed in (Passonneau, 1998). The second most common transition (19%) is Kameyama's Center Establishment, EST (the transition between an utterance without 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, EST and ZEROs, the preferences are roughly as predicted by Brennan *et al.*: the Page test for ordered alternatives ((Siegel and Castellan, 1988), p. 184-188) indicates a chance 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]

	MUSEUM	PHARMA	TOTAL
Continuations Sequences :	10	6	16
Continuation / Retain :	9	3	12
Establishment /Continuation :	17	18	35
Retain Sequences :	5	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 4
Rule 2 statistics considering sequences of transitions.

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 significant, as shown in Table 4. As we'll see again in the Discussion, in our corpus there seems to be a preference for avoiding repetition; this tendency is confirmed by these figures, that 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

23 Similar results were obtained by (Passonneau, 1998).

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 emphasize, 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.

	MUSEUM	PHARMA	TOTAL
Cheap transitions :	76	63	139
Expensive transitions :	263	380	643
Cheap transition pairs :	21	14	35
Expensive transition pairs :	162	234	396

Table 5
Cheap and expensive transitions with the Vanilla instantiation.

These figures 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)–i.e., their CB is the same as their CP.

Salience and Pronominalization The statistics for pronominalization are shown in Table 6. As said above, our corpus contains 217 uses of third person pronouns, 23 demonstratives, and 78 complementizers.[25] In this instantiation we only take R1-pronouns to include 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 antecedent in the same utterance, and 28 (13%) are 'long-distance pronouns' (Hitzeman and Poesio, 1998) whose antecedent is neither in the same nor the previous utterance.

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.

	MUSEUM	PHARMA	TOTAL
Total number of R1-pronouns:	200	95	295
Number of personal pronouns:	144	73	217
Number of complementizers:	56	22	78

Table 6
R1-pronouns in the corpus with the Vanilla instantiation

Table 7 shows that the relation between pronominalization and CB with the Vanilla instantiation is not straightforward: only 55% 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 primarily 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]

	MUSEUM	PHARMA	TOTAL (PERC)
Total number of realizations of CBs:	211	163	374
Total number of CBs realized as R1-pronouns:	138	68	206 (55%)
– 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%)
Complementizers that do not realize CBs:	3	2	5 (6.4%)

Table 7
CBs and pronominalization with the Vanilla instantiation

	MUSEUM	PHARMA	TOTAL
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%)
Gordon - utterances that satisfy:	77	45	122 (44.5%)
Gordon - utterances that violate:	60	92	152 (55.5%)

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

Table 8 analyzes pronominalization in terms of the three versions of Rule 1 we are

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).

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 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 towards the end of texts, when pronouns are sometimes used to realize the product described by the leaflet. E.g., *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 it for longer.

What we seem to observe here is a conflict between the 'global' preference to realize the

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 (27.21% of the total number of utterances) are considered relevant for Rule 1.

29 The problem of intrasentential pronouns in Centering is discussed, e.g., in (Walker, 1989; Tetreault, 2001; Poesio and Stevenson, To appear).

'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 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 72 cases of pronouns that do not refer to the CB, but do not violate Rule 1 fall in two classes: (i) R1-pronouns used in utterances without a CB (the majority), and (ii) 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 do not realize the CB, which is what one would expect on the basis of e.g., (Gundel, Hedberg, and Zacharski, 1993; Passonneau, 1993). This suggests that treating demonstrative pronouns as R1-pronouns would not lead to improvements wrt Rule 1. On the other hand, because there is only 23 of them, this change is 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 are significant. The results reported in the rest of the paper are all obtained with the 'narrow' definition of R1-pronoun that does not include demonstratives.[31]

30 See also (Giouli, 1996; Byron and Stent, 1998).
31 The relation between demonstrative NPs in general and the CB in our corpus is analyzed in detail in (Poesio and Nygren-Modjeska, 2003).

Differences between 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 pharmaceutical domain (73 of 1308 CFs, or 5%, as opposed to 144 of 1731, 8%, for the museum domain), and whereas in the museum domain 40% (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%, 189/430) than in the pharmaceutical one (27.2%, 157/577), and the percentage of utterances with no CB and that are not segment boundaries is much higher in this second domain (55.3%, 319/577) than in the first (39%, 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 is that the layout plays a much more important role in the pharmaceutical domain, providing a different way of achieving coherence. (See Discussion.)

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 paper, we discuss the results with most of the variants quite briefly, and only analyze in some length the instantiation that identifies utterances with sentences; the results are summarized with graphs. The extended report contains a more extensive discussion of some of the variants; the in-

terested readers are also encouraged to try further instantiations in the companion 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.

Treating coordinated VPs *as utterances* Many researchers working on spoken dialogues or NLG assume that each element of a coordinated VP counts as a separate utterance: i.e., 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 of course results in more utterances (1041 vs. 1007) 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 non-significant differences for Rule 1 and Rule 2 (with slightly higher numbers of continuations, and slightly lower of retains).

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, e.g., the infinitival *to*-clause in *John wants to visit Bill*. The results with this instantiation, as well, crucially depend on our grammatical assumptions. With this setting we get of course many more utterances (1267 instead of 1007), 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 non-finite 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 found no significant differences in the number of violations of Rule 1. As for Rule 2, this change

results in significantly fewer NULL transitions (45% instead of 47.9%), and significantly more EST (22.1% instead of 18.8%) and SSH (5.6% instead of 3.7%).

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 1007 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 increase in the number of utterances that satisfy Rule 1 (GJW 95) to 281. The violations to Rule1 are also reduced to 8, 2.8% (not significant). (There are virtually no changes with Rule 2.) Because of these small improvements, in the rest of the paper we always exclude these clauses when discussing the results with finite clauses as utterances; we refer to this instantiation as 'Vanilla-'.

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 are 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. Instead, 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 paper we will continue to count relative clauses as finite clauses.

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

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. HE ...* .

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.

the setting used so far.) When the parameter is set to *(generalized) Suri-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'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 definition satisfy it under Suri'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 Kameyama's and Suri's definition of previous utterance have mostly to do with a type of clause that was only discussed briefly by Kameyama and not at all by Suri and McCoy, relative clauses, as in:

(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' 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 them 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 where the difference does have to do with the way adjuncts

are handled is (7), reproduced again below. 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 of 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 paper 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.

Sentences The setting of **uttdef** with the most dramatic impact on Strong C1 is to identify utterances with sentences. The reasons for this were already 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 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).

	MUSEUM	PHARMA	TOTAL (PERC)
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 (39.7%)
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%)

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

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 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 Encyclopdie, 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 it should hang above the fireplace.

The pronoun *it* in (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, (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

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 goes up quite a bit, from 505 to 669, and that Strong C1 is not verified, even though it would be if only the 505 sentences were considered (the sign test gives $p \leq 0.001$ for Strong C1).

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).

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 figures 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.

	MUSEUM	PHARMA	TOTAL (PERC)
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 (40.9%)

Table 10
Rule 2 statistics with sentences as utterances

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 ten times as many expensive-expensive sequences (191) than cheap-cheap ones (18).

Interim Summary The effect of the changes in the definition of utterance and previous utterance on Strong C1 and Rule 1 are summarized in Fig. 1 and Fig. 2, respectively. As the figures show, most such changes have fairly small effects, even though they 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 non-finite

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

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 drop 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 where '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.

Figure 1
The effect of utterance parameters on Strong C1: a summary

In the rest of the paper 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).

Figure 2
The effect of utterance parameters on Rule 1 (BFP): a summary

4.3 Realization

In this section we discuss the effect of changes in the values of the realization parameters: **realization** and **CF-filter**.

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%) having exactly one CB, and 325 having zero or more than one (33.5%), even Strong C1 is verified by the sign test (+525, -325).[36]

However, allowing for indirect realization has a negative effect on other claims, just

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

like 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' instantiation (1.3%) to 22 (2.3%). This is because by increasing the number of 'persistent entities', we increase the chance of them having an equivalent ranking in the previous utterance. The number of violations of Rule 1 exactly doubles: from 8 with the Suri 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, the instantiations of 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 pronoun that becomes a violation of Rule 1 (GJW 95) if we allow for 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) *One stand* was adapted in the late 1700s or early 1800s century to make it 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%). The facts that there are many more RET than CON, and many more RSH than SSH, mean 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%).

	MUSEUM	PHARMA	TOTAL (PERC)
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%)

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

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

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. 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%).

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+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 segment boundary:	47	55	102 (15.2%)
Utterances without CB :	30	121	151 (22.6%)

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

The number of violations to 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 a 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 does 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% of the total (as opposed to 22% with DS and 27.7% with IF).

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

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*, i.e., 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 should be treated as R1-pronouns. It turns out that doing this actually results in *more* violations of Rule 1; this is because most bridging references do not refer either to the CB of the present utterance or of 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.

38 Walker also observed that in Japanese, zero pronouns–often taken as referring to the CB–are allowed to refer to second person entities (p.c.).

lowing 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 and -242, a significant improvement even over the instantiation with direct realization and PRO2s; with IS, +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% (from 47.6%); for DS, to 30.8% (from 40.8%); for IF, to 18.2% (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 consider second person pronouns as R1-pronouns or not. 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 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 paper 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 website.

Predicative NPs The two alternative views about which entities to realize considered so far both result in an increase in the number of CFs. What if we were to 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 R1 might disappear.

Among the NPs that might be thought not to introduce CFs, an obvious candidate are predicative NPs, i.e., 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 relations 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 listed below. Especially in the second case, it is not clear how else the annotators could have marked the antecedent of *Bjorg*.

39 The difference is significantly worse for all the instantiations not treating PRO2s as CFs; worse, but not significantly so, if PRO2s are treated as CFs.

(19) a. An important artist in making these links has been Yasuki Hiramatsu. His knowledge of metalcraft allows him to push and play against the boundaries of what the material can physically do.

b. Two such jewellers are Toril Bjorg from Norway and Jacqueline Mina from England. It may be unsurprising that Bjorg, as a Scandinavian, should choose silver as her material.

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

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 to Rule 1, although overall the percentage of violations remains pretty small; and it leads to such an increase in the number of RET and RSH, that there are less CON than RET, and less SSH than RSH, that Rule 2 (BFP) is not verified by any of the instantiations with indirect realization we have seen. Treating PRO2s as realizations of CFs, while sufficient to make Strong C1 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

Grammatical Function + Linear Disambiguation We observed above that because grammatical function does not always specify a unique most highly ranked CF, using this ranking function means that some utterances end up with more than one CB, which

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.

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; Gordon, Grosz, and Gillion, 1993), 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 has already been used by Strube and Hahn (1999) to resolve tie-breaks, albeit in conjunction with a different 'basic' ranking function. In this section we evaluate the ranking function obtained by adding linear order to grammatical function, that we call GFTHERELIN.[40] The results with this ranking function are summarized in Table 13.

Instantiation	**Strong C1 (Perc violations)**	**Rule 1 (GJW 95) (Perc violations)**	**Rule 2 (BFP) (Page test, prob. of not being verified)**
DF-predicate	+352,-450 (46.3%) ♠	+291,-11 (3.6%)	.001
DF-predicate+per2	+465,-355 (36.5%)	+403,-15 (3.6%)	.001
DS-predicate	+273,-249 (37.2%) ♠	+259,-14 (5.1%)	.001
DS-predicate+per2	+347,-197 (29.4%)	+325,-22 (6.3%)	.001
IF-predicate	+529,-310 (31.9%)	+463,-18 (3.7%)	1 ♠
IF-predicate+per2	+635,-219 (22.5%)	+325,-22 (3.7%)	.05♢
IS-predicate	+408,-157 (23.5%)	+378,-30 (7.4%)	.05♢
IS-predicate+per2	+469,-113 (16.9%)	+432,-37 (7.4%)	1 ♠

Table 13
Summary of results for Strong C1, Rule 1 (GJW 95) and Rule 2 (BFP) with GFTHERELIN ranking.

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; and the results of the Page test for Rule 2. ♠ indicates that a claim is not verified at the .05 level; ♢ that it's not verified at the .01 level.

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

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

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% 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 R1 ends up not being verified. The results with the DF instantiations aren't significantly worse: e.g., 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 but with simple grammatical function ranking. The number of violations of R1 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 coffers are lined with tortoiseshell and brass or pewter, with secret compartments in the base.

(s8) The coffers are each decorated using techniques known as *premiere partie* marquetry, a pattern of brass 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* become its only CB.

With the IF instantiation, the percentage of violations of Rule 1, 3.7%, is non- significantly 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 also 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 GFTHERELIN, 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 instantiation, we get small but significant increases in CON and RET, and a reduction in RSH. We report the complete percentages for this instantiation, for comparisons with other ranking functions.

	MUSEUM	PHARMA	TOTAL
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%)

Table 14
Transition percentages for IS with GFTHERELIN ranking.

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 are CF realizations, even though Rule 2 is only verified with this instantiation at the .05 level. We will therefore concentrate on this instantiation when making comparisons with the other ranking variants.

Linear Order Among the ranking functions alternative to grammatical function, perhaps the simplest is the one that ranks CFs in the order of 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; 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 where the differences are significant is Rule 2 (BFP): with IS, enough CON become RET, and enough SSH become RSH that Rule 2 is not anymore verified even at the .05 level. (The rule is still verified with the DF and the DS instantiations.)

All in all, these results are not grounds to argue that linear order is a better ranking

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.

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.

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 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 1 more violation than with GFTHERELIN.[43] This doesn't mean that the exact same utterances are violations in both cases; rather, than 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), including 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.

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

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, e.g., Prasad and Strube (2000) for Hindi.

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

As *the corner cupboard* is in object position, it gets higher ranking in s67 than *Count Branicki*, which is in NP modifier position, that–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, therefore the CB of s68; hence the violation. Conversely, (22) is an example in which GFTHERELIN 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, where, 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 the Discussion.

The one claim where STRUBE-HAHN 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 make Rule 2 (BFP) verified at the .01 level with all instantiations considered.[44] The transition percentages with IS and STRUBE-HAHN ranking are in Table 15.

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

	MUSEUM	PHARMA	TOTAL
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%)

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

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

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 with the three versions of ranking we considered. More differences can be found with Rule 2 (BFP), which is not verified by any instantiation with linear order ranking, and only by a few instantiations with GFTHERE-LIN. 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 (BFP) being falsified. Finally, not even these last changes to parameter setting were sufficient to make either Rule 1 (Gordon *et al.*) or Rule 2 (Strube and Hahn) verified.

5 DISCUSSION

We discuss first the effects of different parameter settings; we then analyze the claims of the theory, draw a few theoretical conclusions, and make some suggestions for further work (empirical and theoretical).

5.1 Setting the parameters

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 are set as suggested by Strube and Hahn, and viceversa. 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 GJW 83) are clearly verified. The results concerning Constraint 1 are especially negative. As with this instantiation only 35% of utterances are continuous– i.e., $\text{CF}(\text{U}_n) \cap \text{CF}(\text{U}_{n-1}) \neq \emptyset$ (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 only required 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 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: e.g., 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., $\text{CP}(\text{U}_{n-1})$ predicts $\text{CB}(\text{U}_n)$).

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: i.e., 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 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 a bridging 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, which we called IF and IS, Strong C1 is verified, as well as the two 'basic' versions of R1.

We also found, however, that there is a tradeoff 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 to 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), a claim 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

45 Perhaps the most spectacular demonstration of the tradeoff 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 paper, but can be examined on the companion website.) 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 very few utterances have a CB.

ranking, RET are the second most common transition. As a result, Rule 2 (BFP) is only verified with IS instantiations 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%–is 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 make Rule 2 (BFP) 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 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, had to do with the general issue of which entities should be included in the CF list. Considering second person pronouns realizations of discourse entities is enough to make Strong C1 satisfied; we also found that a number of extensions to the definition of utterance, such as the inclusion of relative clauses, and non-finite clauses, led to much worse results unless reduced relative clauses and non-finite clauses were taken to include traces linking these clauses to the embedding one.

Minimizing violations should not be the overriding goal We already said in Section §3 that we don't think that minimizing violations should be the only factor taken into account when 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 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 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 would need to supplement our studies with psychological experiments reconciling these results with the numerous results indicating the important role played by grammatical function, and 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 reaching conclusions about when the local focus is updated. For one thing, most analysis of discourse structure– e.g., 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 most more likely to be a prosodic phrase of some sort.

5.2 The claims of Centering, revisited

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; Grosz, Joshi, and Weinstein, 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 in general, 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 situation in which a mismatch

between the CB and pronominalization may occur: by having been mentioned in a discourse often, 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]

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 it 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; Joshi and Weinstein, 1981), it says nothing about entity coherence being what ensures local coherence.

Further light on entity coherence is shed by recent work on text planning, particularly by Karamanis (2003), that suggests that when all alternative ways of extracting a text plan from the propositions expressed by texts such as those we are studying are

46 The role of the 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.

considered, the actual ordering found in the texts tends to be in greater agreement with Centering's preferences about entity coherence than most of its alternatives. After extracting the propositions[48] expressed by texts in the museum domain 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 alternative, coherence-inducing device are rhetorical relations. Indeed, the claim that 'entity' coherence needs to be supplemented by 'relational' coherence can already be found in (Kintsch and van Dijk, 1978; 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 domain 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 viceversa utterances that make a

48 More precisely, the lists of CF realized by each utterance with a DF instantiation, representing that utterance's arguments.

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–e.g., 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' which, when looked at closely, turn out to be really 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]

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), that 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, less than one-third of sequence pairs involve the same transition, and even less 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 all together only 28% of sequence pairs only involve transitions considered by Grosz *et al.*. Keeping in mind that Rule 2 (GJW 95) only applies 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

50 As said earlier, an earlier version of Kibble's proposal was also tested; the results can be viewed on the companion web site.

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, that 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)

5.3 Theoretical Consequences

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

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 the theory by identifying aspects that need to be made more precise. Our study raised a number of questions about the

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.

definitions of the concepts used in Centering not previously mentioned 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 indicate that if PRO2s are not considered realizations of CF, or we treat them as R1-pronouns, we find many more violations of Strong C1 and R1, respectively (although both claims are still verified). We also saw that the results concerning Constraint 1 and Rule 1 depend on whether reduced relative clauses and non-finite VPs are assumed to contain traces, and whether these traces were assumed to be R1-pronouns or not. More in general, we identified the need for a clear definition of 'R1-pronoun': i.e., 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, e.g., 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 only relevant 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 is the previous utterance 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* newspaper:

(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.

Separating entity coherence from CB *uniqueness* Starting with (Brennan, Friedman, and Pollard, 1987) and, more recently, (Beaver, 2004; 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, 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 argued throughout the paper 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; Karamanis, 2001), we will call this first half of Constraint 1 (ENTITY) CONTINUITY:

(Entity) Continuity: $\text{CF}(\text{U}_{i-1}) \cap \text{CF}(\text{U}_i) \neq \emptyset$

Weak C1 is CB Uniqueness, whereas Strong C1 is CB uniqueness plus Continuity.

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 into 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 links must exist. 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}) ∩ CF(U_i) ≠ ∅;
2. or there is a rhetorical relation **RR** such that **RR**(U_{i-1},U_i);[52]
3. or U_{i-1} and U_i are temporally coherent in the sense, e.g., of

(Kameyama, Passonneau, and Poesio, 1993);

4. ... (other)

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 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.

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 so, 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 or 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 of them occupies a particularly salient position either in

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.

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 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 papers such as (Gordon, Grosz, and Gillion, 1993) suggest that when two animate entities are considered, only one tend 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 is perhaps most clearly associated with the work of Givon (1983). Within the Centering literature, abandoning the claim that we called 'CB Uniqueness' has been suggested by Gundel (1998), and, more radically, in work such as (Strube, 1998; Gordon and Hendrick, 1999; Tetreault, 2001), where the whole notion of CB is abandoned.

As seen in Section §2, the primary motivation for CB uniqueness are complexity-theoretic arguments: inference in monadic logics is less expensive than with normal logics (Joshi and Kuhn, 1979; Joshi and Weinstein, 1981). Grosz and colleagues's linguistic evidence for CB uniqueness are constrasts 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 only observed 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) ev-

idence 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 of 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; Gundel, 1998). In both cases, we would need a separate theoretical account of RNP effects. More empirical evidence is needed on this issue.[53]

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 a half of CBs are realized as R1-pronouns. It is also the case that CBs are hardly ever continued for more than 2-3 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 2/3 of all transition sequences involve two different transitions. In fact, we hypothesize that the Repeated Name Penalty observed by Gordon *et al.* might be an instance of this more general phenomenon.

ACKNOWLEDGMENTS

Special thanks to Nikiforos Karamanis, Alistair Knott, Mark Liberman, Ruslan Mitkov, Jon Oberlander, Tim Rakow, and the other members of the GNOME project: Kees van Deemter, Renate Henschel, Rodger Kibble, Jamie Pearson, and Donia Scott. We also wish to thank James Allen, Jennifer Arnold, Steve Bird, Susan Brennan, Donna Byron,

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 non-empty. In particular situations, that 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.

Herb Clark, George Ferguson, Jeanette Gundel, Aravind Joshi, Eleni Miltsakaki, Rashmi Prasad, Ellen Prince, Len Schubert, Joel Tetreault, Lyn Walker, and audiences at the ACL 2000, the University of Pennsylvania, the University of Rochester, CLUK, and the University of Wolverhampton for comments and suggestions. The corpus was annotated by Debbie De Jongh, Ben Donaldson, Marisa Flecha-Garcia, Camilla Fraser, Michael Green, Shane Montague, Carol Rennie, and Claire Thomson, together with the authors. A substantial part of this work, including the creation of the corpus, was supported by the EPSRC project GNOME, GR/L51126/01. Massimo Poesio was supported during parts of this project by an EPSRC Advanced Fellowship. Barbara Di Eugenio is supported in part by NSF grant INT 9996195, in part by NATO grant CRG 9731157. Janet Hitzeman was in part supported by the EPSRC project SOLE, GR/L50341.

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.

Baker, Collin F., Charles J. Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proc. of the 36th ACL*.

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 *Proc. of the 25th ACL*, pages 155–162, June.

Byron, Donna and Amanda Stent. 1998. A preliminary model of Centering in dialog. In *Proc. of the 36th ACL*.

Caramazza, Alfonso, E. Grober, C. Garvey, and J. 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 *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford.

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

Cooreman, Ann and Tony Sanford. 1996. Focus and syntactic subordination in discourse. Research Paper 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, chapter 4, pages 55–70.

Dale, Robert. 1992. *Generating Referring Expressions*. The 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, chapter 7, pages 115–138.

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

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

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, University of Edinburgh, Linguistics Department.

Givon, Talmy, editor. 1983. *Topic continuity in discourse : a quantitative cross-language study*. J. Benjamins.

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.

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

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.

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

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, 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 *Proc. 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 *Proc. of the International Conference on Spoken Language Processing (ICSLP98)*,, Australia.

Hitzeman, Janet and Massimo Poesio. 1998. Long-distance pronominalisation and global focus. In *Proc. 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 8th Annual Meeting of the Cognitive Science Society*, 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 in Discourse*. Oxford University Press, pages 199–226.

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

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

Joshi, Aravind K. and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure–centering. In *Proc. IJCAI*, pages 435–439.

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 *Proc. ACL-86*, 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, chapter 6, pages 89–112.

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

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

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

Karamanis, Nikiforos. 2003. *Entity coherence for descriptive text structuring*. Ph.D. thesis, University of Edinburgh, Informatics.

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

Kibble, Rodger. 2000. A Reformulation of Rule 2 of Centering Theory. Technical report, University of Brighton, ITRI. GNOME project internal deliverable.

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 *Proc. of INLG*, 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.

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, May.

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 *Proc. of CLIN*.

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. Presented at the Berlin workshop on Topics in Discourse, September.

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

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, chapter 17, pages 327–358.

Passonneau, Rebecca. and Diane Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*.

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, chapter 7, pages 179–227.

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

Pearson, Jamie, Rosemary Stevenson, and Massimo Poesio. 2001a. The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In M. Poesio, editor, *Proc. 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 *Proc. of the 2nd LREC*, pages 211–218, Athens, May.

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

Poesio, Massimo. 2004. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. 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, *Proc. 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, *Proc. of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

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

Poesio, Massimo and Rosemary Stevenson. To appear. *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, June.

Prasad, Rashmi and Michael Strube. 2000. Discourse salience and pronoun resolution in Hindi. In *Penn Working Papers in Linguistics*, volume 6, 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*. John Benjamins, pages 295–325.

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

Rambow, Owen. 1993. Pragmatics aspects of scrambling and topicalization in german. In *Proc.*

of the Workshop on Centering Theory in Naturally-Occurring Discourse, Philadelphia. Institute for Research in Cognitive Science (IRCS).

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

Scott, Donia, Richard Power, and Roger Evans. 1998. Generation as a solution to its own problem. In *Proc. of the 9th 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, MIT.

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

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.

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.

Strube, Michael. 1998. Never look back: An alternative to centering. In *Proc. 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, U. 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 in Discourse*. Oxford University Press, 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. Squib.

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proc. ACL-89*, pages 251–261, Vancouver, CA, June.

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

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

Walker, Marilyn A., A. K. Joshi, and E. 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*. Clarendon Press / Oxford, chapter 1, pages 1–28.

Walker, Marilyn A., A. K. Joshi, and E. F. Prince, editors. 1998b. *Centering Theory in Discourse*. Clarendon Press / Oxford.

Webber, Bonnie L. 1978. A formal approach to discourse anaphora. Report 3761, BBN, Cambridge, MA, May.