Natural Language Engineering Group
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom

Centering: a parametric theory and its instantiations

Massimo Poesio
University of Essex
poesio@essex.ac.uk

Rosemary Stevenson
University of Durham
Rosemary.Stevenson@durham.ac.uk

Barbara di Eugenio
The University of Illinois at Chicago
bdieugen@cs.uic.edu

J. Hitzeman
The MITRE Corporation
hitz@mitre.org

NLE Technical Note TN-02-01
(Revised May 2004)

CS Technical Report CSM-369

This Report is an extended version of a paper that will appear in *Computational Linguistics*, 30(3), 2004

University of Essex, Department of Computer Science, NLE Group

Other Technical Notes and theses from the Natural Language Engineering and Web Applications group are available electronically at

`http://cswww.essex.ac.uk/Research/nle.htm`

Centering: a parametric theory and its instantiations

Abstract

Centering Theory has proven an extremely useful conceptual framework for theorizing about local coherence and salience. Theoretical concepts such as 'utterance', 'previous utterance', 'realization', and 'ranking' have served as the basis for theories of a number of discourse phenomena across several languages, such as, e.g., (zero) pronominalization. But because its proponents intended to formulate preferences that would be cross-linguistically valid and would abstract away from specific generation or interpretation algorithms, these concepts were intentionally left underspecified in the formulation of the theory; as a result, the claims the theory makes have only ever been tested by fixing upon a particular way of instantiating these PARAMETERS– e.g., by assuming that ranking is based on grammatical function. This leaves open the question of whether all the particular parameters instantiations proposed in a particular study were optimal. Furthermore, none of the previous corpus-based studies included a test of the reliability of the annotation.

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'–that each utterance has exactly one CB– 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. Rule 2 is also sensitive to parameter instantiation, although more robust than Constraint 1. 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– cannot be the only criterion which makes discourses locally coherent; itmust be supplemented at least with an account of relational coherence.

1 MOTIVATIONS

Centering Theory (Joshi and Weinstein 1981; Grosz et al. 1983, 1995; Walker et al. 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, and a key difference from earlier theories of focusing more directly concerned with anaphora resolution such as Sidner's (1979), is that it is best viewed as a *linguistic* theory-and a very abstract one- than a computational one. By this we mean, first of all, 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). And second, that the theory aims to specify a 'conceptual vocabulary' of discourse notions which can be used to make such claims; this vocabulary is meant to play a role analogous to that played in syntax by notions such as 'command' or 'specifier'.

The result is a very different theory from those one usually finds in Computational Linguistics. Particularly disturbing to many computational linguists is the fact that in central papers such as (Grosz et al. 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 et al. 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 et al. 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, often by the same authors: e.g., different definitions of CB can be found in (Grosz et al. 1983, 1995; Gordon et al. 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. This situation is not unusual for a linguistic theory–consider, for example, the situation encountered in syntax, where very many alternative definitions of the same notion of 'command' have been proposed–but makes Centering rather different from the theories typically encountered in computational linguistics.

This lack of algorithmic detail shouldn't lead to the conclusion that Centering is merely an attempt at mapping out a field of research, without making any specific claims. On the contrary, the theory does make very strong claims, discussed in a moment. It does mean, though, that two 'instantiations' of the Centering framework may both be compatible with the framework's central claims, yet make different predictions, just as different ways of specifying 'command' may result in different predictions about binding violations or about the possible distinct scopal readings of a sentence (for examples of the latter, see, e.g., (Reinhart 1983; May 1977, 1985; Szabolcsi 1997)). This freedom to 'fill the gaps' has proven inspirational for researchers, who have devoted themselves to provide such definitions, for specific languages or in general; so much so that the conceptual framework provided by the earlier papers in Centering Theory has been the basis for most work on local salience in computational linguistics, and even in psychology, in the last ten years. But in part because of the underspecification, in part because of the existence of so many competing instantiations, many researchers have expressed doubts about the empirical status of the theory: i.e., about the extent to which its claims are supported by empirical evidence, and how they are affected by the way the parameters are specified. In order to be meaningful, this comparison should be done using the same data

set.

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 (a task not always made easy by the terminology of 'Constraints' and 'Rules'); 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 et al. 1986; Gordon et al. 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, as this fact may be extremely useful for the developers of NLP systems, whether from an interpretation or a generation perspective.)

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. 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. In Section §5 we re-examine a few linguistic claims that make use of notions from Centering at the light of the results in Section §4. In the following section we briefly report a second series of experiments investigating the impact of rhetorical structure. Our results are discussed in Section §7.

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 et al. 1995; Walker et al. 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. This makes Centering a theory of pronominalization as well, under the assumption that the most salient entities are most likely to be pronominalized (Grosz et al. 1995; Gundel et al. 1993).

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; Ariel 1990; Gundel et al. 1993) and is supported by much psychological evidence (Hudson et al. 1986; Gernsbacher and Hargreaves 1988; Gordon et al. 1993; Stevenson et al. 1994).

These claims about coherence and salience are linked by two further hypotheses: that the identity of the CB is crucially determined by the entities' ranking, and that the CB is most likely to be realized as a pronoun. This assumption that a 'main entity' or 'topic' or 'focus' is the preferred interpretation of pronouns is commonly found in theories in the psychological (e.g., (Sanford and Garrod 1981)), computational (Sidner 1979) and linguistic literature (Ariel 1990; Gundel et al. 1993) and is motivated by evidence such as the contrast between examples (1) and (2).[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.

[1] Entity-based theories of coherence are so-called by contrast with RELATION-CENTERED theories of coherence, such as those developed in (Hobbs 1979; Reichman 1985; 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; further discussion of 'entity-centered' vs. 'relation-centered' notions of coherence is in (Knott et al. 2001; Stevenson et al. 2000).)

[2] (1) is Grosz et al. (1995)'s (15), p. 215; these examples are also discussed by Kehler (1997).

(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]

(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.

As evidence for the hypothesis that ranking affects coherence, Grosz *et al.* produce examples like (4) and (5):[4]

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

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

According to Grosz *et al.*, although both discourses express the same information, in the first discourse the discourse entity *John* is the most highly ranked in all utterances; this 'packaging' of the information (Vallduvi 1990) conveys the impression that all utterances are 'about' the same discourse entity, *John*, which makes the discourse highly coherent. In (5), by contrast, utterance (b) and (d) are constructed in such a way that *the store* is ranked more highly than *John*; although in fact *John* is still the center of every utterance, these continuous changes in ranking suggest that the discourse does not have a clear center. The result is that the reader finds this second text less coherent.

[3]These are examples (6) through (10), Grosz, Joshi and Weinstein, 1995, p. 211-212.
[4]These are examples (1) and (2) from (Grosz et al. 1995).

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).[5] 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 et al. (1987) formalized the relationship between utterances and CFs by means of one of their so-called 'Constraints':[6]

Constraint 2: Every element of the list of forward centers for U, CF(U,DS), must be REALIZED in U.

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; Anderson et al. 1983; Givon 1983; Reichman 1985; Vallduvi 1990; Gundel et al. 1993) and plays a central role in the theory's claims about both coherence and salience. Although in the original paper on Centering (Grosz et al. 1983) the CB was only characterized in intuitive terms, most subsequent work within the framework has been based on the definition below of the CB of utterance U_i in terms of ranking (Grosz et al. 1995), called 'Constraint 3' by Brennan et al. (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.(This is in contrast with, for example, the view expressed by Sidner in her dissertation, where determining the (discourse) focus involved complex computations also taking into account, for example, which entities had been referred to, and for how long.)

Alternative definitions of CB Gordon et al. (1993) propose to replace the definition of CB in Constraint 3 with an operational one, i.e., a test that can be used to identify the CB. More specifically, they propose to identify the CB with the entity which is subject to the REPEATED NAME PENALTY

[5]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; Groenendijk and Stokhof 1991).

[6]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.

(RNP) discussed later in this section (a slower reading time whenever a full NP is used instead of a pronoun). Gordon *et al.*'s experiments suggest that RNP effects only occur with subjects referring to a subject or first mention antecedent; as a result, they propose that the CB should be identified with the subject 'if possible'. We interpret this claim as meaning that the CB should be identified with the subject whenever the subject does refer to a discourse entity realized in subject or first-mention position in the previous utterance.

This new definition of CB is in conflict with that based on Constraint 3; indeed, the experiments reported by Gordon *et al.* (especially experiment 2) show that NPs that satisfy Constraint 3 (according to Gordon et al's own definition of ranking) are not always subject to the RNP. Example (5) illustrates these differences: whereas the discourse entity *John* is the CB of all utterances between (5b) and (5d) in this example if we use Constraint 3 for our definition, none of these utterances has a CB according to the definition provided by Gordon *et al.*.[7]

A second operational definition of the CB was proposed by Passonneau (1993) on the basis of her analysis of the uses of *it* and *that* in dialogues. Passonneau notices how difficult it is to identify the CB on semantic / pragmatic grounds, and, like Gordon *et al.*, proposed to use preferred pronominalization patterns to identify it, using however the new term 'Local Center' to denote this operationally defined entity. In particular, she proposed a specific linguistic context as one of Local Center Establishment:

> **Local Center Establishment Rule**:
>
> **A. Recognizing a Local Center:** Two utterances U1 and U2 that are adjacent in their segment establish a discourse entity E as a local center only if U1 contains a third person, singular, non-demonstrative pronoun N1 referring to E, U2 contains a co-specifying third person, singular, non-demonstrative pronoun N2, and N1 and N2 are both subjects or non-subjects, in that order of preference.
>
> **B. Generating a Local Center:** To establish a discourse entity E as a local center in a pair of adjacent utterances U1 and U2, use a third person, singular, non-demonstrative pronoun to refer to E in both utterances. Both pronouns should be subjects or non-subjects, in that order of preference.

Obviously this definition is even more restrictive than Gordon *et al.*'s; the extent to which this is the case will become clearer when discussing the results obtained with this version.

Transitions The hypothesis that discourses are perceived to be (locally) more coherent when successive utterances are packaged in such a way as to be perceived as being 'about' a unique discourse entity (see example (4)) is formalized in Centering as a preference for certain ways of updating the local focus. This preference is formulated 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 et al. (1995) distinguish between three types of transitions, depending on whether the backward looking center of U_{i-1} is maintained or not in U_i, and on whether CB(U_i) is also the most highly ranked entity (CP) of U_i:

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

[7]Notice that Gordon *et al.* are *not* claiming that the subject is the most highly ranked CF: this latter claim concerns the identity of the CB in the *following* utterance, whereas the claim by Gordon *et al.* concerns the position of the CB in the *current* utterance.

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 et al. 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 et al. 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.

We will now discuss each of these claims and some of their variants proposed in the literature.

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 et al. 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 (henceforth, Rule 1 (GJW 95), 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 et al. (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.

Their observations of the Repeated Name Penalty effect discussed earlier led Gordon et al. (1993) to propose 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)).[8] This evidence led them to propose , in addition to the more restrictive definition of CB discussed above, 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.)

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. Some of these alternative were motivated by the goal of achieving a better account of local (entity-based) coherence, by finding a definition that would reflect the actual preferences observed in texts (e.g., (Strube and Hahn 1999)). Other proposals were motivated by evidence about the distribution of NP forms: in particular, the distinction between 'weak' forms such as pronouns in English or zeros in Italian and Japanese, thought to be preferred for expressing continuations, and 'strong' forms, thought to be used to indicate shifts (Di Eugenio 1998; Turan 1998). We considered a number of such proposals.

The version of Rule 2 presented in (Grosz et al. 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 on weak vs strong forms of pronouns 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 et al. 1987; Walker et al. 1994, 1998a). The version of Rule 2 proposed by Brennan *et al.* (and slightly revised by Walker *et al.*) is as follows:

Rule 2 (Single transitions): Transition states are ordered. The CON transition is preferred to the RET transition, which is preferred to the SMOOTH-SHIFT transition (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:

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

[8] Specifically, 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.

	$CB(U_n) = CB(U_{n-1})$ 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

an utterance without one, such as the first utterance of a segment. Walker et al. (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.[9] 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 (what we'll call here the 'NULL' transition).

Strube and Hahn (1999), like Grosz et al. (1995), claim that inferential load is evaluated across sequences (pairs, in fact) of transitions, but argue for a different way of evaluating the inferential load of utterances. In their view, classifications of transitions such as those above do not reflect what should be one of the crucial claims of the theory: that the CP of one utterance predicts the CB of the next. In order to formalize their view, they propose a different classification scheme, based on the distinction between CHEAP and EXPENSIVE transitions ((Strube and Hahn 1999), 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 the backward-looking center of the current utterance is not correctly predicted by the preferred center of the previous utterance, i.e., 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.

Finally, Kibble (2001) argues that we should view the two dimensions of classification used by Brennan et al - whether the CB of the current utterance is the same as the CB of the previous utterance, and whether the CB and the CP of the current utterance coincide - as reflecting respectively the degree to which the current utterance is coherent with the previous utterance, and the degree to which it makes the CB most salient. He then argues that while it's the case that, given the principles inspiring Centering, utterances that satisfy both criteria - CONs - should be most preferred, and utterances that satisfy neither - RSHs - most dispreferred, there isn't any obvious *a priori* reason why coherence should be preferred to salience, i.e., RET to SSH, as argued by Brennan *et al.*.

As a result, Kibble proposes to replace the single Rule 2 of previous versions of Centering with a collection of principles stating preferences; and that these principles may conflict with each other. Kibble proposed two versions of Rule 2, in an earlier version of his paper (Kibble 2000) and a final one in Kibble (2001). The version of Kibble's formulation of Rule 2 that we tested, in (Kibble 2000), is as follows:

Rule 2 (Kibble): Continuity: prefer transitions such that $CF(U_n) \cap CF(U_{n-1}) \neq \emptyset$.

[9] This proposal is reminiscent of Sidner's idea that the first utterance only introduces an 'Expected Discourse Focus', to be confirmed later.

Salience: prefer transitions such that CB(U_n) = CP(U_n).

Cheapness: prefer transitions such that CB(U_n) = CP(U_{n-1}).

Cohesion: prefer transitions such that CB(U_n) = CB(U_{n-1}).

Kibble doesn't commit to a particular way of resolving conflicts between these principles, but mentions that one way would be to treat all principles as ranked equally and to prefer the interpretation (or to produce the utterance) that satisfies the largest number of them, as done in (Kibble and Power 2000); a second way would be to establish preferences among them and choose the interpretation that violates the weakest constraints, as done in Optimality Theory.[10]

Coherence and Salience In combination, Constraint 1, Rule 1, and Rule 2 express what is perhaps the most distinguishing feature of the Grosz and Sidner's theory of the attentional state as articulated, say, in (Grosz and Sidner 1986): that coherence and salience are strongly tied, both at the global level, whose building units are discourse segments, and at the local level, whose building blocks are utterances. At the global level, according to Grosz and Sidner, perceiving a text as coherent depends on the ability to establish relations between the intentions expressed by the segments, and salience is 'parasitic on the intentional structure.' At the local level, the data structure whose values determine whether a text is perceived as being coherent, the CB, also determines which entity is most salient in the sense of being most likely to be pronominalized. An additional (implicit) assumption is that coherence within a segment is (largely) *entity* coherence, whereas global coherence is mainly of the intentional / rhetorical / relational sort (Kintsch and van Dijk 1978; Stevenson et al. 2000).[11] Finally, it is important to stress that these claims are meant to indicate *preferences* rather than hard-and-fast constraints.

> ...the most fundamental claim of Centering Theory [is] that to the extent a discourse adheres to Centering constraints, its coherence will increase and the inference load placed upon the hearer will decrease. (Grosz et al. 1995, p.210)

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.[12]

[10] See (Karamanis 2003) for further discussion and an evaluation of the effect of these two ways of resolving conflicts.

[11] See (Knott et al. 2001) for an argument that in certain genres at least, global coherence may depend on entity coherence.

[12] For more details, and for a discussion of the motivations behind these proposals, see (Poesio and Stevenson to appear).

Utterance and Previous Utterance In the early Centering papers, utterances were implicitly identified with sentences. Kameyama (1998), however, argued that such identification makes the number of potential antecedents of anaphoric expressions much greater than if they were resolved clause by clause. Furthermore, she noted that this identification leads to problems with multiclausal sentences: for example, grammatical function ranking becomes difficult to compute, as a sentence may have more than one subject. Kameyama proposed that the local focus 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 (6) into utterances as follows, and to treat each of these utterances, including subordinate clauses such as (u2) or (u5), as an update:

(6) (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 et al. (2000) confirmed that CFs introduced in main clauses are significantly more likely to be subsequently mentioned than CFs introduced in complement clauses, which supports Kameyama's claim that complements should be treated as embedded. 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.[13] 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 et al. (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),[14] 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:

(7) (u1) John walked towards *the house*. (u2) The door 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

[13] Hurewitz (1998) introduces a distinction between unrestricted relative clauses with clause-final head nouns, treated as utterances, and other relatives, treated as part of the main clause.

[14] Associative references are one type of BRIDGING REFERENCE (Clark 1977).

reference *the door* (see, e.g., the discussion in (Grosz et al. 1995; Walker et al. 1998b)). Clearly, the computation of the CB is affected by which entities are considered to be 'realized' in an utterance: in (7), 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) focus to be realized in an utterance in these cases, and the issue is often mentioned in discussions of Centering.

A related issue is whether empty realizations, or *traces*, should count as realizations of an entity. Many theories of grammar hypothesize that morphologically null elements occur in the syntactic structure underlying a variety of constructions, including control constructions as in (8a), reduced relatives as in (8b), and even coordinated VPs as in (8c):

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

If, e.g., the coordinated VP in (8c) 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 (9), 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.[15]

(9) (u1) You should not use PRODUCT-Z
(u2) if you are pregnant of breast-feeding.
(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 et al. 1994, 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, 1986) and (Grosz et al. 1986) have assumed that GRAMMATICAL FUNCTION plays the main role in determining the order among forward looking centers, at least for English. Specifically, (Grosz et al. 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 et al. 1986)). Slightly different ranking functions based on grammatical function where proposed by Brennan et al. (1987) (who made a further distinction between objects

[15] According to Walker et al. (1998a), in the original version of (Grosz et al. 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.

and indirect objects), by Walker et al. (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 et al. 1986; Gordon et al. 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 Corbett and Chang (1983) and Gernsbacher and Hargreaves (1988) suggest that order of mention affects recall from memory, and in particular, 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 et al. (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.[16] Order of mention also plays a role in their ranking: within each category, the entities realized earlier in the sentence are ranked more highly.

More formally, Strube and Hahn characterize ranking as a partial order relation $\prec$, defined as follows:

1. If x belongs to OLD and y belongs to MED, x $\prec$ y
2. If x belongs to OLD and y belongs to NEW, x $\prec$ y
3. If x belongs to MED and y belongs to NEW, x $\prec$ y
4. If x and y belong to the same set (OLD, MED, or NEW) and x precedes y, x $\prec$ y
5. Otherwise, x and y are unordered.

Finally, Sidner's 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 et al. 1994; Pearson et al. 2001). In particular, there is evidence that with certain verbs, the normal preference for subjects to rank higher than their objects is reversed; and in transfer sentences, THEMEs are ranked more highly than GOALs, which in turn are ranked more highly than SOURCEs, although these preferences are modified by other factors such as order of mention, the type of connective, and animacy (Stevenson et al. 1994, 2000; Pearson et al. 2001).[17]

[16]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.

[17]Pearson et al. (2001) found that ANIMACY plays an even stronger role than thematic roles or order of mention: between an animate entity and an inanimate one, the animate is always most salient (= most likely to be

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 et al. 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.

Segmentation and the relationship between global focus and local focus According to Grosz and Sidner (1986), Centering is only meant to capture preferences about coherence and salience within discourse segments. A proper evaluation of the claims of the theory would require therefore a corpus annotated for intentional structure. However, neither Grosz et al. (1995) nor Grosz and Sidner (1986) provide a specification of discourse intentions (Discourse Segment Purposes) explicit enough that can be used to identify the intentional structure of texts; as a result, only preliminary attempts at annotating texts according to Grosz and Sidner's theory have been made. In the central sections of this paper we only discuss the results obtained when segmenting the text according to a few heuristic methods, including that proposed by Walker (1989). In Section §6 we present further results obtained using a distinct corpus, independently annotated according to Relational Discourse Analysis, a technique for analyzing real texts inspired by Grosz and Sidner's proposals (Moser and Moore 1996b), and from which a Grosz and Sidner-like segmentation was extracted using methods proposed in (Poesio and Di Eugenio 2001). We also report in Section §5 results concerning the question of whether the distinction between the two levels results in linguistic differences, i.e., whether pronouns are preferred for references within the local focus whereas definite descriptions or full NPs are used for global focus reference (see, e.g., "a particular claim of Centering Theory is that the resource demands of this inference process are affected by the *form of expression* of the noun phrase .. " (Grosz et al. 1995, p.208) as well as (Gundel et al. 1993)).

One type of linguistic usage that blends these boundaries are long-distance pronouns (Fox 1987; Hitzeman and Poesio 1998; Hahn and Strube 1997). Hitzeman and Poesio found that while the antecedents of long distance pronouns are always within the stack, as suggested by (Grosz 1977; Fox 1987), not all discourse entities could serve as antecedents; there was an additional requirement that the antecedent had to have been a CB (similar findings were reported by (Iida 1998; Brennan 1998)). We collected statistics about long-distance pronouns and their correlation with CB and CP.

We will not be concerned here with issues raised by studies that challenge the theoretical model proposed by Grosz and Sidner, e.g., by arguing that the stack is not an appropriate model of the global focus (Walker 1996, 1998) or that global coherence may be based on entities rather than intentions (Knott et al. 2001). For a discussion of some of these questions, see (Poesio and Di Eugenio 2001).

pronominalized), irrespective of thematic role or order of mention. This would support the claims about the role of animacy made by, e.g., (Sidner 1979; Dahl and Fraurud 1996). We did not study the effect of animacy on ranking in this study.

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 et al. 1994; Di Eugenio 1998; Turan 1998) and several papers in (Walker et al. 1998b)). The claims about pronominalization made in Centering have been applied to develop algorithms for both anaphora resolution (Brennan et al. 1987; Strube and Hahn 1999; Tetreault 2001) and for sentence planning (Dale 1992; Hitzeman et al. 1997; Henschel et al. 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 et al. 1986; Hudson-D'Zmura and Tanenhaus 1998). And we already mentioned that Gordon et al. (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. (This may come as a surprise especially to those familiar with the psychological literature on Centering, in which some papers seem to assume that the parameters of Centering are completely and uniquely specified–typically, according to what we will call below the 'Vanilla' instantiation of the theory.) 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 annotation methods, and how the annotation was used.

3.1 Evaluating the Claims of Centering against a Corpus

A preliminary question we had to address is what are in fact the main claims of the theory. As discussed in Section §2, of the seven claims mentioned in (Grosz et al. 1995), Constraint 1, Rule 1, and Rule 2 are the ones that can actually be verified using a corpus; we concentrated on these. (The development of a 'conceptual vocabulary' for theories of local coherence and local salience is of course a significant contribution, but the same notions are assumed by all instantiations.) Because several variants of these three claims have been proposed, we evaluated a few of these variants as well. We also report the results obtained by considering Gordon *et al.*'s and Passonneau's definitions of CB instead of Constraint 3.

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.[18] 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 always 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.[19] 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 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 would like to stress 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.

[18] Beaver (2004) argues–correctly, in our opinion–that in one of the best-known pronoun resolution algorithms based on Centering, that proposed by Brennan et al. (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.

[19] This point is especially important from an NLG perspective: see, e.g., (Karamanis 2003). We will return on this issue in the Discussion.

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.[20] The pharmaceutical subcorpus is a selection of leaflets providing the patients with legally mandatory information about their medicine.[21] The GNOME corpus also includes tutorial dialogues from the Sherlock corpus collected at the University of Pittsburgh (Lesgold et al. 1992; Di Eugenio et al. 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.

The data used for this study have two characteristics that make them of particular interest. First of all, they cover genres not previously considered in studies on Centering, and more similar to those that 'real' NLP applications have to contend with.[22] Secondly, and most-importantly, these texts are strongly entity-centered (see, e.g., (Knott et al. 2001) for an analysis of the museum data), so the hypotheses about coherence formulated in Centering are likely to play an important part in the way these texts are constructed.

3.3 Annotation

The previous corpus-based investigations of Centering Theory we are aware of (Walker 1989; Passonneau 1993, 1998; Byron and Stent 1998; Di Eugenio 1998; Hurewitz 1998; Kameyama 1998; Strube and Hahn 1999) were all carried out by a single annotator annotating her/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 the full annotation manual, see the companion web site.

How can Constraint 1, Rule 1 and Rule 2 be evaluated 'in a general way', when their definitions rely on notions that different authors specify in different ways? Any attempt at annotating a corpus for 'utterances', or their CBs, is bound to force the annotators to adopt a specific setting of these basic concepts; the problem is even worse with psychological experiments. Because of this,

[20] 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).

[21] 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 et al. 1998).

[22] The genres normally used to study Centering Theory are ones thought to be more 'natural,' such as narratives, spoken dialogues, or newspaper articles. This makes sense from a scientific point of view, but does raises the question of whether the preferences about coherence and salience expressed by Centering Theory might not be overridden by other factors in other genres.

previous psychological studies and corpus investigations of the theory have generally focused on a specific variant of the theory (Byron and Stent 1998; Di Eugenio 1998; Gordon et al. 1993; Gordon and Scearce 1995; Gordon and Chan 1995; Gordon et al. 1999; Hudson-D'Zmura and Tanenhaus 1998; Kameyama 1998; Passonneau 1993; Walker 1989; Walker et al. 1994).[23] Yet, different ways of specifying the parameters of Centering could result in very different theories, at least in principle; and, most importantly, these studies cannot test whether a different combination of parameter settings from those proposed in the literature might lead to better results. The only way around this problem seems to consider many different ways of setting the parameters, compare them, find if one or more of these instantiations are significantly better than the others (the 'best' way being the one that results in the fewest violations of Constraint 1, Rule 1, and Rule 2), and use these versions to assess the claims of the theory.

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).[24] 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. Examples of clauses, verbal and non-verbal parentheticals, preposed PPs, and coordinated VPs marked as units follows; the parentheses indicate unit boundaries. (Sentence boundaries are not indicated.)[25]

(10) a. **clausal unit**: (They were founded in 1903 by Josef Hoffmann and Koloman Moser)

b. **clausal unit with non-verbal parenthetical**: (It's made in the shape of a real object (– a violin))

[23] A few studies compare two instantiations of the theory: e.g., two or more algorithms for anaphora resolution (Strube and Hahn 1999; Tetreault 2001), two views of ranking (Strube and Hahn 1999; Prasad and Strube 2000) or competing theories of transitions (Passonneau 1998; Strube and Hahn 1999; Gordon et al. 1999).

[24] Of course, many studies compare two or three instantiations of the theory: e.g., two or more algorithms for anaphora resolution (Strube and Hahn 1999; Tetreault 2001), two views of ranking (Strube and Hahn 1999; Prasad and Strube 2000) or competing theories of transitions (Passonneau 1998; Strube and Hahn 1999; Gordon et al. 1999).

[25] 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).

c. **clausal unit with preposed** PP **and embedded relative clauses**: ((With the development of heraldry in the later Middle Ages in Europe as a means of identification), all (who were entitled (to bear arms)) wore signet-rings (engraved with their armorial bearings))

d. **clausal unit with non-finite complement clause and coordinated** VP: (The center of the narrow body swells (to allow for the pendulum's swing), (and has a viewing hole to observe the movement))

As example (10d) above illustrates, subordinate units such as clausal complements and relative clauses are enclosed within the superordinate unit in the annotation. Subordinate units also include adjunct clauses headed by connectives such as *before*, *after*, *because* and clauses in subject position.

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. The possible values for this attribute are main, relative, such-as, appositive, parenthetical, paren-rel, paren-app, paren-main, subject, complement, adjunct,coord-vp,preposed-pp, listitem, cleft, title, disc-marker.
- verbed: whether the unit contains a verb or not.
- finite: for verbed units, whether the verb is finite or not.
- subject: for verbed units, whether they have a full subject, an empty subject (expletive, as in *there* sentences), or no subject (e.g., for infinitival clauses).

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:

Attribute	κ **Value**
utype	.76
verbed	.9
finite	.81
subject	.86

The main problems encountered in marking up units were to identify complements, to distinguish clausal adjuncts from prepositional phrases, and how to mark up coordinated units. The main problem with complements was to distinguish non-finite complements of verbs such as *want* from the non-finite part of verbal complexes containing modal auxiliaries such as *get*, *let*, *make*, and *have*:

(11) a. (I would like (to be able to travel))

b. (I let him do his homework)

One problem that proved fairly difficult to handle (and which, in fact, we couldn't entirely solve) was clausal coordination. The problem was to preserve enough structure to be able to compute the previous utterance, while preserving some basic intuitions about what constitutes a clause (roughly, that by and large clauses were texts spans marked either by the presence of a semantically isolated verb or by punctuation / layout) which are essential for annotators and are needed to specify the values of attributes. This was relatively easy to do when two main clauses were coordinated, since the

embedding sentence could be used to preserve the information that the two units occurred at the same level; coordinated main clauses were marked as in (12a). However, it wasn't completely obvious what to do in the case of coordination within a subordinate clause, as in (12b). Because there weren't many such cases, rather than using the unit element with a special value for utype as we did for coordinated NPs (which meant specifying all sorts of special values for attributes) we used a markup element called unit-coordination to maintain the structure, and then marked up each clause separately, as shown in (12c) (where the unit-coordination is marked with square brackets).

(12) a. (The Getty museum's microscope still works,) (and the case is fitted with a drawer filled with the necessary attachments).

b. (If you have any questions or are not sure about anything, ask your doctor or your pharmacist)

c. ((If [(you have any questions) or (you are not sure about anything)]), ask your doctor or your pharmacist)

Our genres raised two issues that, as far as we know, have not been previously discussed in the Centering literature. One such problem is what to do with layout elements such as titles and list elements, which can clearly serve as the first introduction of a CF and to move the CB. One example of title unit is unit (u1) in (13).

(13) (u1) Side effects

Side effects may occur when PRODUCT-Y is applied to large parts of the body,

We addressed this problem by marking up these layout elements as units, as in (14), but using the special value title of the 'unit type' attribute utype (see above) so that we could test whether it was better to treat them as utterances or not.

(14) (u1)<unit id="u1" utype="title">Side effects</unit>

(u2) <s stype="decl"><unit> Side effects may occur <unit>when PRODUCT-Y is applied to large parts of the body, ... </unit> ... </unit> ... </s> ...

The elements of text *not* marked up as units include: NPs, post-verbal and post-nominal PPs, non-verbal NP modifiers, coordinated VPs in case the second conjunct did not have arguments (15a), and quoted parts of text, when they are not reported speech (15b).

(15) a. (The oestradiol and norethisterone acetate are plant derived and synthetically produced)

b. (The inscription 'CHNETOC BASHLHKOC CPATHARHC')

In total, the texts used for the main study contain 505 sentences and more than 1,000 units, including 900 finite clauses. (Notice that the number of utterances depends on how utterances are defined.)

Concerning attributes, one problem we had (especially with the pharmaceutical texts) was instructions in the imperative form, as in (16). The problem was addressed by marking up finiteness, rather than tensedness as originally proposed by (Kameyama 1998), since imperative clauses are considered finite although they are not tensed.

(16) (u1) Gently rub the correct amount into the skin (u2) until it has all disappeared.

The most difficult attribute to mark was utype, and our main problem was to distinguish between relative clauses and parentheticals, since it's not always easy to tell whether a relative clause is restrictive or non-restrictive (see also (Cheng et al. 2001)). In the end, we adopted rules purely based on syntax

(the presence or absence of a comma or other bracketing device). (See also (Quirk and Greenbaum 1973).) The two tables summarize the distribution of utterances in our corpus.[26]

Total number of utterances:	1578
Values of FINITE:	
finite-yes	916
finite-no	304
no-finite	358
Values of VERBED:	
verbed-yes	1218

Values of UTYPE:	
main	628
complement	162
relative	136
adjunct	94
preposed-adjunct	62
preposed-pp	47
coord-vp	49
subject	3
parenthetical	98
appositive	12
paren-app	62
paren-rel	38
paren-main	5
such-as	16
title	69
listitem	86
captionitem	2
disc-marker	2
unsure-utype	7

NPs Our instructions for identifying NP markables derive from those proposed in the MATE scheme for annotating anaphoric relations (Poesio et al. 1999), in turn derived from DRAMA (Passonneau 1997) and MUC-7 (Chinchor and Sundheim 1995). As in the case of units, the main problem with marking up NPs was coordination. Our approach was to use a separate ⟨ne⟩ element to mark up the coordinated NP, with type (cat) value coord-np. We only used a coord-np element if two determiners were present, as in *((your doctor) and (your pharmacist))*. This approach was chosen because it limited the number of spurious coordinations introduced (in cases such as *this is an interesting and well-known example of early Byzantine jewellery*), but has the limitation that only one ⟨ne⟩ is marked in cases such as *Your doctor or pharmacist*. The distribution of NPs in the corpus used for this study according to their type is as follows:

[26]Because of all these issues, although we tried a couple of automatic parsers at the beginning of our annotation effort, we didn't really feel we could use them to do the markup (more precisely, we found it faster for a trained annotator to mark up the units by hand rather than to correct the problems with the output of the parser). Because of the rapid improvements in parsing technology, soon enough it might be worth reconsidering this decision.

Total number of NPs:	3345
Type (= Values of CAT):	
Pronouns:	
pers-pro (1st, 2nd and 3rd)	324
poss-pro	208
this-pro	21
q-pro (e.g., pronominal *any*, *each*)	18
num-ana (e.g., *I want three*)	7
refl-pro	3
null-ana	3
that-pro	2
Definite NPs:	
the-np	554
the-pn	71
pn	320
poss-np	250
this-np	91
that-np	4

Indefinite NPs:	
bare-np	745
a-np	269
num-np (e.g., *three cars*)	71
meas-np (e.g., *three pounds of X*)	23
another-np	11
Other:	
q-np	117
coord-np	114
gerund	44
complementizer	43
wh-pro	8
wh-np	5
such-np	4
free-rel	5
unsure-cat	10

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 complete list of values is: a-np, another-np, q-np, num-np, meas-np, that-np, this-np, such-np, wh-np, poss-np, bare-np, pn, the-pn (for definites that are really disguised proper names, such as *the Beatles*), the-np, pers-pro, poss-pro, refl-pro, rec-pro, q-pro, wh-pro, this-pro, that-pro, num-ana (for 'numerical anaphors' such as *one* in *I want one*), null-ana, gerund (for nominalized present participles such as *veneering furniture* in *the practice of veneering furniture*), coord-np, and free-rel (for 'free relatives' such as *what you need most* in *what you need most is a good rest*)).

- 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 et al. 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:

Attribute	κ **Value**
cat	.9
gen	.89
gf	.85
num	.84
per	.9

Other attributes of NPs we could reliably annotate include ani (whether the object denoted is animate or inanimate), count (whether the NP is countable or not), deix (whether the object is a visual deictic reference or not), generic (whether the NP denotes generically or not), lftype (whether the NP is the realization of a discourse entity, a quantifier, or a predicate), loeb (its functionality or lack of it under the scheme proposed by (Loebner 1987)), onto (its ontological status - denoting a concrete object, an event, a time interval, or an abstract entity), its structure (whether it denotes a set or an atom) (Poesio 2000). 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 et al. 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 et al. 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 <u>its</u> 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 et al. 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. The few cases where information about semantic relations of this type may be needed are cases in which this information may detemine the choice of CB, and these can usually be dealt with by marking multiple semantic relations between entities in the present and in subsequent utterances, rather than between entities in the same clause. For example, in cases like:[27]

(17) (u1) PRODUCT-Z$_i$ is one of a group of medicines called corticosteroids$_j$.
(u2) <u>These</u>$_k$ can be used to relieve the symptoms of hay fever or rhinitis.

where it is important to know that a semantic relation exists between *PRODUCT-Z* and *a group of medicines called corticosteroids* because this information will result in (u2) being treated as a continuation rather than as a smooth-shift (if indirect realization is allowed), the necessary information can be marked by annotating *two* semantic relations between entity$_k$ and entities in (u1): an identity

[27] Thanks to an anonymous reviewer for this example.

relation with entity$_j$ (the corticosteroids), and a part-of relation with entity$_i$. Also, our texts do not include any case of bound anaphora,[28] 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.[29] (See (18).)

(18)
```
<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 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 are a subset of those proposed in the 'extended relations' version of the MATE scheme and include set membership (ELEMENT), subset (SUBSET), and 'generalized possession' (POSS), which includes part-of relations as well as ownership relationsand their inverses. 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. They were told to mark all identity relations, but to mark associative relations

1. only with entities last mentioned in a different unit;
2. never more than one associative relation for each anapohoric expression;
3. always choosing the closest entity;

Furthermore, we specified preferences: for example, in *Francois, the Dauphin*, the embedding NP would be chosen as an antecedent of subsequent anaphoric references, 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

[28] Strictly speaking, in dynamic theories of anaphora *all* anaphoric expressions are 'bound' by their antecedent. We only refer here to cases of anaphoric expressions bound by non-existential quantifiers.

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

the other.[30] 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).

Segmentation According to Grosz and Sidner (1986), Centering is only meant to capture preferences about coherence and salience within discourse segments. A proper evaluation of the claims of the theory would require therefore a corpus in which discourse segments have been identified.[31] Unfortunately, discourse segments are difficult to identify reliably (Passonneau and Litman 1993; Marcu et al. 1999), and Grosz and Sidner (1986) do not provide a specification of discourse intentions (Discourse Segment Purposes) 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. Our own preliminary experiments didn't give good results, either.

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 addition to Walker's heuristic, we considered (i) not segmenting texts at all, (ii) using each main section of a text as rough segments, and (iii) treating every paragraph as a separate segment. In the museum domain, we identified as separate sections each description of an object or an artist; in the pharmaceutical domain, we considered as a separate section each of the legally required subsections of a leaflet ('What this medicine is for', 'What's in your medicine', etc.). In Section §6 we present further results obtained using a distinct corpus, independently annotated according to Relational Discourse Analysis, a technique for analyzing real texts inspired by Grosz and Sidner's proposals (Moser and Moore 1996b), and from which a Grosz and Sidner-like segmentation was extracted using methods proposed in (Poesio and Di Eugenio 2001). We also report in Section §5 results concerning the question of whether the distinction between the two levels results in linguistic differences, i.e., whether pronouns are preferred for references within the local focus whereas definite descriptions or full NPs are used for global focus reference (see, e.g., "a particular claim of Centering Theory is that the resource demands of this inference process are affected by the *form of expression* of the noun phrase .. " (Grosz et al. 1995, p.208) as well as (Gundel et al. 1993)).

One type of linguistic usage that blends these boundaries are long-distance pronouns (Fox 1987; Hitzeman and Poesio 1998; Hahn and Strube 1997). Hitzeman and Poesio found that while the an-

[30] 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. The only method we could see of ensuring an equal number of 'categories' for anaphoric expressions was to include all CFs in the segment, which in some cases means considering more than 100 'categories'. With an average 3 CFs per finite clause, and 5 CFs per sentence, the number of potential antecedents for an anaphoric expression ranges between 15 and 25, even if we just consider the previous 5 utterances.

[31] This has been contested in work such as (Di Eugenio 1998; Walker 1998; Strube and Hahn 1999), on the grounds that it is not entirely clear whether local structure is meant to be entirely embedded within global structure - i.e., whether the theory's claims are intended to operate purely within segments - or if in fact the two structures are independent of each other, with local transitions possibly operating across segment boundaries (see, e.g., (Walker 1998)).

tecedents of long distance pronouns are always within the stack, as suggested by (Grosz 1977; Fox 1987), not all discourse entities could serve as antecedents; there was an additional requirement that the antecedent had to have been a CB (similar findings were reported by (Iida 1998; Brennan 1998)). We collected statistics about long-distance pronouns and their correlation with CB and CP.

3.4 Automatic computation of Centering information

Perl scripts working off the annotated corpus automatically compute 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: Grosz Joshi and Weinstein's, Gordon *et al.*'s, or Passonneau's.

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

prev(ious utterance): treat adjunct clauses Kameyama-style or Suri-style.[32]

neverutt: the clauses that should never be considered as utterances, even if finite or verbed.

realizes: Controls realization. Only allow direct realization, or indirect realization via bridging references as well.

cfselect: 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*).

ranking: rank CFs according to grammatical function, linear order, a combination of the two as in (Gordon et al. 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*).

segment(ation): identify segments using Walker's heuristics, or with paragraphs, sections, or whole texts.

prepadj: whether the computation of the previous utterance for preposed adjunct clauses (e.g., *if*-clauses, as in *if X, Y*) should follow the linear order, or the subordination order.

bridges_policy: whether implicit anaphoric elements such as those occurring in traces should be counted as pronouns for the purposes of Rule 1 or not.[33]

[32] In fact, the instantiation we call here 'Kameyama' treats *all* types of clauses other than complement clauses –including, e.g., relative clauses–as not embedded, whereas the instantiation we call 'Suri' treats all such clauses as embedded, including clauses that Suri and McCoy didn't consider themselves; so these names should be taken with a grain of salt. One case in which these differences matter is the discourse *This brooch is made of titanium, which is one of the refractory metals. It was made by Anne-Marie Shillitoe, an Edinburgh jeweller, in 1991.*. The 'Kameyama' instantiation assigns the relative clause *which is one of the refractory metals* as previous utterance to the clause *It was made by ...*; whereas the 'Suri' instantiation treats the relative clause as embedded. We return to this issue below.

[33] Relative pronouns (implicit and explicit) were only counted as pronouns if not doing so would lead to a violation of Rule 1.

The scripts consider utterance u_i a violation of (Strong) Constraint 1 if either no entity realized in the previous utterance u_{i-1} is realized in u_i, or if two entities with the same rank in u_{i-1} are both realized in u_i. Utterances are classified into transitions for the purposes of (several versions of) Rule 2 using the tables discussed in Section §2. The only aspect that does need some explanation is how the violations of the three versions of Rule 1 are computed. The basic logic is very simple: for each utterance *u*

1. If *u* has no CB, it is ignored (i.e., utterances without a CB are never considered violations);
2. Else, if CB(*u*) is realized *at least once* as a R1-pronoun, count *u* as a verification (+) for all three versions of Rule 1 that we are considering;
3. Else, carry out *all three* of the following actions:
   (a) Count *u* as a violation (-) of Gordon *et al.*'s version of Rule 1 (always pronominalize the CB)
   (b) If CB(*u*) = CB(*u-1*), count *u* as a violation of the version of Rule 1 from (Grosz et al. 1983), else as a +;
   (c) If at least one entity other than the CB is realized as a R1-pronoun, count *u* as a violation of the version of Rule 1 from (Grosz et al. 1995) (pronominalize the CB if anything else is), else as a +.

The one additional complication 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 et al. 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, and finally of different definitions of CB. 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 et al. 1995) and (Brennan et al. 1987), and incorporating additional suggestions from (Kameyama 1998), (Gordon et al. 1999), and (Walker et al. 1998a). The Vanilla instantiation is based on the definition of CB from (Grosz et al. 1995), and uses grammatical function for ranking, as proposed there and in (Brennan et al. 1987) (also incorporating the proposals concerning ranking in complex NPs from (Gordon et al. 1999)). 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.[34] 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 et al. 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.[35] 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.[36]

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 henceforth write +833, -11 to indicate verifiers

[34] 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.

[35] This is a major difference from our previous work (Poesio et al. 2000).

[36] For those intending to replicate the results using the website, the complete setting we used was as follows:

```
-uttdef finite -realizes direct -ranking gf -poss_np gordon -coord_np
gordon -prev kameyama -prepadj linearize -cfselect "per2" -alwaysutt
"" -neverutt "coord-vp" -segment walker -nosegboundary title -cbdef bfp
-output file -prodef narrow -bridges_policy noempty
```

The last test was run in February 2003. The results for this configuration, and with the rest, may slightly change when new annotation problems are found, or problems with the script are corrected; after several years of testing however such changes have become quite rare.

	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.

	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.

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% by the sign test.[37] 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 (19) 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.[38]

(19) (u1) These "egg vases" are of exceptional quality:

(u2) basketwork bases support egg-shaped bodies

(u3) and bundles of straw form the handles,

(u4) while small eggs resting in straw nests serve as the finial for each lid.

(u5) Each vase is decorated with inlaid decoration: ...

Clearly, there are two ways of 'fixing' this problem with the Vanilla instantiation. 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

[37]Furthermore, the figure of 346 utterances verifying Strong C1 includes 71 relative clauses whose only reference to entities in the embedding clause is their complementizer or a trace.

[38]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.

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 et al. 1998b), p. 3),[39] 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 (20), where we show the XML markup so that the attributes of elements are visible:

(20)
```
<unit finite='finite-yes' id='u227'>
   <ne id='ne546' gf='subj'>The drawing of
      <ne id='ne547' gf='np-compl'>the corner cupboard</ne></ne>
   <unit finite='no-finite' id='u228'>,   or more probably
      <ne id='ne548' gf='no-gf'> an engraving of
         <ne id='ne549' gf='np-compl'>  it </ne></ne>
   </unit>,
   must have caught
   <ne id='ne550' gf='obj'>
      <ne id='ne551' gf='gen'>Branicki's </ne> attention</ne>
</unit>
<unit id="u229" finite="finite-yes">
   <ne gf="subj" id="ne552">Dubois</ne> was commissioned through
      <ne gf="adjunct" id="ne553">  a Warsaw dealer </ne>
      <unit id="u230" finite="finite-no"> to construct
         <ne gf="obj" id="ne554">  the cabinet </ne>
        for<ne gf="adjunct" id="ne555">the Polish aristocrat</ne>
      </unit>
</unit>
<ante current='ne554' rel='ident'><anchor antecedent='ne549'>
</anchor></ante>
<ante current='ne555' rel='ident'><anchor antecedent='ne551'>
</anchor></ante>
```

In this example, two discourse entities introduced in utterance u227 are realized in utterance u229:[40] *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, as far as we know; 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 results concerning Rule 2 allow us to explore some of the issues about coherence raised by the results about Constraint 1 in greater detail. The statistics about transitions relevant for Brennan *et al.*'s version of Rule 2 are shown in Table 3.

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

[40] Neither u228 nor u230 are treated as utterances as they are not finite.

	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 one also without a CB. (Examples include u3, u4, and u5 in (19).) 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 (19).) CON, RET, SSH, and RSH follow. 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. [41]

	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.[42] As we'll see again in the

[41] Similar results were obtained by (Passonneau 1998).

[42] We should add that we used the most favourable way of counting sequences–each pair of repeated transi-

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 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.[43] Ignoring the 225 segment boundary utterances,[44] 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.

We devised the following method to evaluate Kibble's proposal (the original version in (Kibble 2000)). We counted the total number of utterances verifying one of Kibble's four constraints; we also computed a 'Kibble score' for each utterance, defined as the number of constraints satisfied by that utterance. With the Vanilla instantiation, the average Kibble score[45] comes to about 1.05 - i.e., each utterance satisfies about one of the four constraints. The figures are as follows:

	MUSEUM	PHARMA	TOTAL
Continuous transitions :	195	162	357
Salient transitions :	109	110	219
Cheap transitions :	76	63	139
Cohesive transitions :	59	49	108
Average 'Kibble Score' :	1.29	.87	1.05

Salience and Pronominalization The statistics for pronominalization are shown in Table 6. As said above, our corpus contains 217 uses of third person pronouns (*he*, *she*, *it*, *they*, and their morphological variants), 23 demonstratives, and 78 complementizers.[46] In this instantiation we only take R1-pronouns to include personal pronouns and complementizers, for a total of 295 R1-pronouns. If

tions was counted as a sequence, which means that three CON in a row count as two sequences.

[43]Similar results were found for dialogues by Byron and Stent (1998).

[44]We will ignore them in the rest of the paper, also when considering Kibble's version.

[45]Defined as $\frac{Continuous+Salient+Cheap+Cohesive}{Utttotal-SegBoundary}$.

[46]We will use the term complementizer to indicate relative pronouns and relative traces.

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. The corpus contains 63 pronoun-pronoun chains (cases in which the antecedent of a pronoun is itself realized as a pronoun).

	MUSEUM	PHARMA	TOTAL
Total number of R1-pronouns:	200	95	295
Number of personal pronouns:	144	73	217
Number of complementizers:	56	22	78
Number of demonstrative pronouns:	7	16	23
Utterances with a subject:	386	216	602
Number of personal pronouns in subject position:	61	34	95
Number of demonstrative pronouns in subject position:	5	11	16

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

The next table,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[47] 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.[48]

	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 realized as demonstrative pronouns:	3	1	4 (1%)
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%)
Demonstrative pronouns that do not realize CBs:	4	15	19 (82.6%)

Table 7: CBs and pronominalization with the Vanilla instantiation

Table 8 analyzes pronominalization in terms of the three versions of Rule 1 we are considering.[49] 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 et al. (1993) to the less restrictive definition of CB given by Constraint 3–is not verified: in fact, 55% of utterances violate it.

[47] Even though only 357 utterances have a CB with this instantiation, a CB may be realized more than once in an utterance.

[48] Earlier versions of these findings led to the development of the pronominalization algorithm in (Henschel et al. 2000).

[49] 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.

	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.

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 (21), CB(u3), *PRODUCT-X* is realized as a proper noun, whereas a possessive pronoun is used to refer intrasententially to *the baby*.[50]

(21)
(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.

(22) (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 'main character' of a discourse as a pronoun, and the 'local' preference to pronominalize the locally most salient entity, as identified by the CB.[51] 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. The following example from the museum texts makes a similar point, in a more compact way:

(23)
(u1) Before 1666 Boulle was awarded the title of master cabinetmaker;
(u2) in 1672 the king granted him the royal privilege of lodging in the Palais du Louvre.
(u3) In the same year, he achieved the title of cabinetmaker and sculptor to Louis XIV, king of France.

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

[51]See also (Giouli 1996; Byron and Stent 1998).

The CB of u3, *Louis XIV*, is realized using a proper name, presumably because it occurs in a reference to an official title; the pronoun *he* is used to realize *Boulle*, which, while the 'main character' of this discourse in the sense of Garrod and Sanford (1994) (and the 'discourse focus' in the sense of Sidner), is not the CB of u3.[52]

We saw in Table 7 that although there are only 9 violations of Rule 1 from (Grosz et al. 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):

(24) (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 et al. 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.[53] The results reported in the rest of the paper are all obtained with the 'narrow' definition of R1-pronoun that does not include demonstratives. An interesting effect of the broader definition of R1-pronoun is a 50% increase in the number of long-distance pronouns, from 28 to 39. The overall results for the different versions of Rule 1 including demonstratives among R1-pronouns are summarized in the following table.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	130	133	263
GJW 95 - utterances that violate:	7	4	11
Gordon - utterances that satisfy:	80	46	126
Gordon - utterances that violate:	57	91	148
GJW 83 - utterances that satisfy:	119	105	224
GJW 83 - utterances that violate:	18	32	50

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

[52]This example is complicated by the issue of whether the reference to Louis XIV in (u3) is a genuine use or a mere mention; but we believe that the example would still work even if instead of utterance (u3), we had (u3′): In the same year, he became cabinetmaker and sculptor for Louis XIV.

[53]To test this configuration, use `wide` instead of `narrow` as the value of the `prodef` parameter –i.e., the parameter controlled by the "definition of pronouns used" on the website.

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. (We'll return on the issue of alternative ways of ensuring a text is coherent in the Discussion.)

4.2 Varying the utterance parameters

We now begin to explore alternative parameter settings. 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. (This position is especially natural in grammatical theories, such as Systemic Functional Grammar, in which coordinated VPs are viewed as sentences with an empty subject). 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,[54] 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). The complete figures for Constraint 1 are as follows:

	MUSEUM	PHARMA	TOTAL (PERC)
Number of times at least one CF(Un) is realized in Un+1:	214	178	392 (37.7%)
Utterances that satisfy Constraint 1 (have exactly one CB) :	207	166	373 (35.8%)
Utterances that do not satisfy Con 1 but are segment boundary:	67	96	163 (15.7%)
Utterances with zero CBs :	167	326	493 (47.4%)
Utterances with more than one CB :	7	5	12 (1.15%)

whereas those for Rule 1 are:

	MUSEUM	PHARMA	TOTAL (PERC)
Utterances considered:	140	137	277
GJW 95 - utterances that satisfy:	132	135	267 (96.4%)
GJW 95 - utterances that violate:	8	2	10 (3.6%)
Gordon - utterances that satisfy:	77	44	121 (43.7%)
Gordon - utterances that violate:	63	93	156 (56.3%)
GJW 83 - utterances that satisfy:	116	104	220 (79.4%)
GJW 83 - utterances that violate:	24	33	57 (20.6%)

[54] To test this configuration, do not specify any value for the `neverutt` parameter–in the website, remove `coord-vp` from the units never to be treated as utterances.

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, although the actual number is slightly lower. 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%).[55]

Treating titles and other layout elements as utterances The evaluation script treats as an utterance every unit which contains NPs and is not embedded in any other unit, irrespective of whether it is finite or a clause, because otherwise these NPs would not belong to any utterance. This feature of the script makes the results for Constraint 1 reported so far significantly better than they would be if we were only considering finite clauses or clauses as utterances, because in this case a large number of titles and other layout units would not be treated as utterances. When only finite clauses are considered, there are more violations of both Constraint 1 and Rule 1, although the difference is only significant only in the case of Strong C1. This is even more true of the case discussed below when utterances are identified with sentences. Titles are treated as utterances in the instantiations studied in the rest of the paper, even when they are not finite clauses or sentences.[56]

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*).[57] 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 97.2%. 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-'.

[55]This configuration can be tested using `verbed` instead of `finite` as the value of the `uttdef` parameter. The version of the script running on the website incorporates the 'trace' mechanism.

[56]This is another difference from the scripts used in (Poesio et al. 2000), in which titles were not automatically treated as utterances.

[57]To test this configuration, specify `coord-vp empty-subject subject parenthetical disc-marker` as the value of the `neverutt` parameter (on the web site, list these types of units among those never to be treated as utterances).

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. Our annotation scheme does not include the information necessary to test the proposal in (Hurewitz 1998) which makes a distinction between different types of relative clauses. In addition, we considered treating relative clauses as adjuncts (i.e., as not embedded) and treating them as complements (embedded).[58] The figures reported so far were obtained by treating relative clauses as utterances, and as akin to adjuncts.[59] Not treating relative clauses as separate utterances[60] 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 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.[61] 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 (25), the previous utterance for (u3) is (u2). (This was 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).

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

Using Suri's definition of previous utterance[62] 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,

[58]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 ...* .

[59]The reader should remember 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.

[60]This configuration can be tested by adding `relative` to the value of the `neverutt` parameter.

[61]Things would be very different if relative clauses were not always assumed to contain a link to the embedding clause. In this case, merging would mean greatly improved results, as reported in (Poesio et al. 2000).

[62]This is done by using `suri` as the value of the `prev` parameter. The figures discussed here were obtained by making just this one change to the 'Vanilla-' configuration.

but not significant improvements for Rule 2(more EST and CON, fewer ZERO). 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). Complete figures for Constraint 1 are as follows.

	MUSEUM	PHARMA	TOTAL (PERC)
Number of times at least one CF(Un) is realized in Un+1:	198	170	368 (37.9%)
Utterances that satisfy Constraint 1 (have exactly one CB):	191	164	355 (36.5%)
Utterances that do not satisfy Con 1 but are segment boundary:	67	92	159 (16.4%)
Utterances with zero CBs :	140	305	445 (45.8%)
Utterances with more than one CB :	7	6	13

The overall figures concerning violations of the different versions of Constraint 1 and Rule 1 with Suri's definition of previous utterance, and the probabilities that these principles are falsified according to the sign test, are as follows:[63]

Principle	Plus	Minus	p
CONSTRAINT 1 (STRONG)	355	457	$p = 1.000$
CONSTRAINT 1 (WEAK)	800	13	$p = 0.000$
RULE 1 (GJW 95)	287	8	$p = 0.000$
RULE 1 (GORDON)	132	163	$p = 0.969$
RULE 1 (GJW 83)	243	52	$p = 0.000$

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:

(26) (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 (9), 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.

(9) (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.

[63]The reader may have noticed that 295 utterances are considered for the evaluation of Rule 1, rather than 281 for the Vanilla- instantiation. This is again because of the way the algorithm for counting violations of Rule 1 works.

(27) (u1) Do not use scissors

(u2) as you may damage the patch inside.

(u3) Take out the patch.

We should also point that complement clauses do not always behave as embedded, as assumed by both Kameyama and Suri and McCoy. In the following example, treating (u2) as embedded means that neither *Louis XIV* nor *the cabinet* are in the CFs list when (u3) is encountered.

(28) (u1) The fleurs-de-lis on the top two drawers indicate
(u2) that the cabinet was made for Louis XIV. (u3) As it does not appear in inventories of his possessions, it may have served as a royal gift.

Treating relative clauses and all types of adjuncts as embedded also leads to better results concerning Rule 2: fewer ZERO transitions, slightly more Center Establishments and Center Continuations, more SSH, more cheap transitions, fewer expensive ones, and a better 'Kibble Score' (1.15 instead of 1.08). The differences between the 'generalized Suri' instantiation and the 'generalized Kameyama' are much less if we don't treat relative clauses as utterances, but for Rule 2, unlike Constraint 1, generalized Suri still behaves slightly better. 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 (9) and (27), 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.[64] The reasons for this were already illustrated with (19): 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.[65] 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).[66]

	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.

[64]The results in this section were obtained by setting `uttdef` to the value `s`, and leaving everything else unchanged.

[65]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).

[66]This is the case even though many titles are excluded by the count as they are treated as segment boundaries.

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:

(29) (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 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),[67] although Rule 1 (Gordon *et al.*) still isn't (+97, -167). It is also interesting to note 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 complete statistics about pronominalization for the instantiation identifying utterances with sentences are as follows:

	MUSEUM	PHARMA	TOTAL
Total number of realizations of CBs:	180	158	338
Total number of CBs realized as R1-pronouns:	88	41	129
CBs realized as personal pronouns:	88	41	129
CBs realized as relative pronouns:	0	0	0
CBs realized as demonstrative pronouns:	4	2	6
CBs NOT realized as R1-pronouns:	92	117	209
Total number of R1-pronouns that do not realize CBs:	53	24	77
Personal pronouns that do not realize CBs:	53	24	77
Demonstrative pronouns that do not realize CBs:	3	11	14

Whereas the numbers of violations and verifications of the various versions of Rule 1 are as follows:

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	119	133	252 (95.5%)
GJW 95 - utterances that violate:	7	5	12 (4.5%)
Gordon - utterances that satisfy:	62	35	97 (36.7%)
Gordon - utterances that violate:	64	103	167 (63.3%)
GJW 83 - utterances that satisfy:	106	103	208 (79.2%)
GJW 83 - utterances that violate:	20	36	56 (20.8%)

The results for Rule 2 are not that different from those obtained with finite clauses, but we do observe more continuations and fewer NULLs. The 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.

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

	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(except that sequences of NULL transitions are still the most common). As for the version of Rule 2 proposed by Strube and Hahn, identifying utterances with sentences reduces the number of expensive transitions; but there still ten times as many expensive-expensive sequences (191) than cheap-cheap ones (18).

	MUSEUM	PHARMA	TOTAL
Cheap transitions :	54	44	98
Expensive transitions :	126	220	346
Cheap transition pairs :	11	7	18
Expensive transition pairs :	58	133	191

And finally, the Kibble score goes up with this instantiation, to 1.4.

	MUSEUM	PHARMA	TOTAL
Continuous transitions :	131	147	278
Salient transitions :	54	87	141
Cheap transitions :	54	44	98
Cohesive transitions :	50	56	106
Average 'Kibble Score' :	1.61	1.27	1.4

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 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.

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).

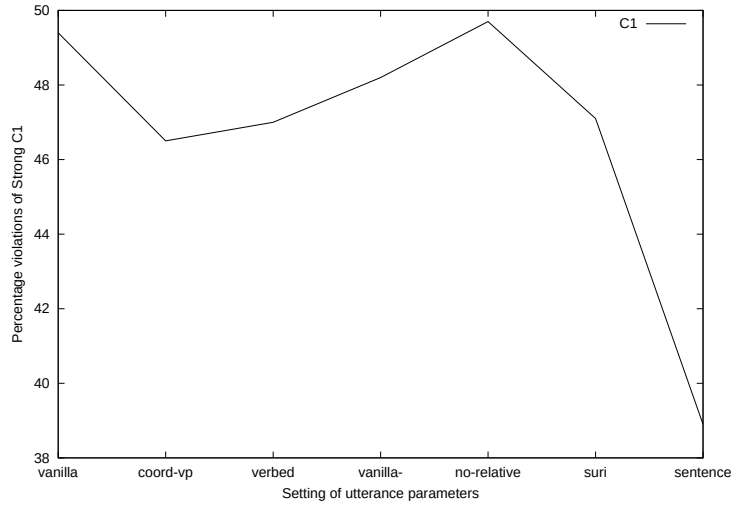

Figure 1: The effect of utterance parameters on Strong C1: a summary

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 (19) 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,[68] 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).[69] The complete figures for this instantiation are as follows.

	MUSEUM	PHARMA	TOTAL (PERC)
Number of times at least one CF(Un) is realized in Un+1 :	299	248	547 (56.3%)
Utterances that satisfy Constraint 1 (have exactly one CB) :	290	235	525 (54.0%)
Utterances that do not satisfy Con 1 but are segment boundary:	48	74	122 (12.6%)
Utterances with zero CBs :	58	245	303 (31.2%)
Utterances with more than one CB :	9	13	22 (2.3%)

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

[68]This configuration can be tested by changing the value of the `realizes` parameter to `indirect`. The results in this section were obtained by leaving all other parameters as in the 'Suri' configuration.

[69]The number of utterances is obviously not affected by changes in the realization parameters.

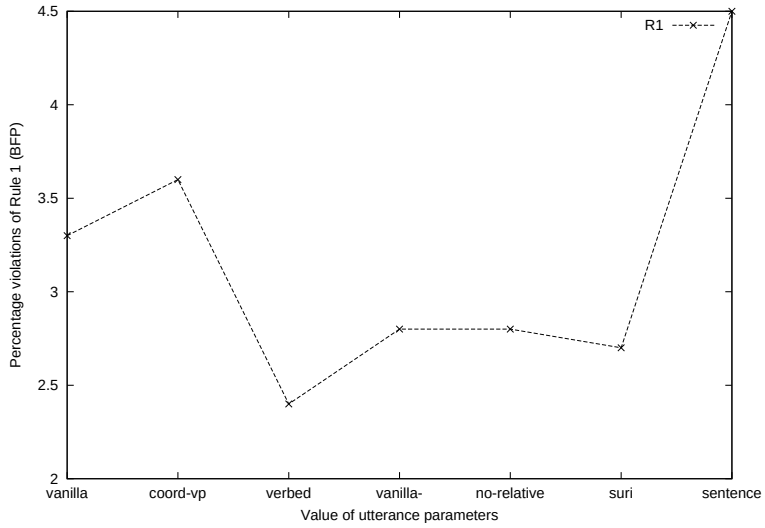

Figure 2: The effect of utterance parameters on Rule 1 (BFP): a summary

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). The overall statistics for pronominalization with this instantiation and u=f are shown in the following table.

	MUSEUM	PHARMA	TOTAL (PERC)
Total number of realizations of CBs:	222	174	396
Total number of CBs realized as R1-pronouns:	135	74	209 (52.8%)
CBs realized as personal pronouns:	95	55	150 (37.9%)
CBs realized as relative pronouns:	40	19	59 (14.9%)
CBs realized as demonstrative pronouns:	3	1	4 (1%)
CBs NOT realized as R1-pronouns:	87	100	187 (47.2%)
Total number of R1-pronouns that do not realize CBs:	46	15	61 (21.7%)
Personal pronouns that do not realize CBs:	43	13	56 (25.8%)
Relative pronouns that do not realize CBs:	3	2	5 (7.8%)
Demonstrative pronouns that do not realize CBs:	4	15	19 (82.6%)

Whereas the figures for validity and violations of the different versions of Rule 1 are as follows:

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	237	214	451 (96.6%)
GJW 95 - utterances that violate:	14	2	16 (3.4%)
Gordon - utterances that satisfy:	84	52	136 (29.1%)
Gordon - utterances that violate:	167	164	331 (70.9%)
GJW 83 - utterances that satisfy:	172	146	318 (68.1%)
GJW 83 - utterances that violate:	79	70	149 (31.9%)

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 (30). 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 is not pronominalized: only *one stand* is.

(30) (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*.

(Of course, this pronoun would not count as a violation if the non-finite clause containing *it* were counted as a separate utterance. Also, it may be thought that the CB *should* be treated as implicitly realized–we'll see below why this is not a good idea.)

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 (30): 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%, as opposed to 98/444, 22% for the u=s instantiation).The Kibble score increases as well, from 1.15 to 1.6 (vs. 1.4 for the u=s instantiation).

	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

In what follows, 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.[70] With this instantiation (henceforth, IS) 390 utterances out of 669 (58.3%) satisfy Strong C1, and 177 (26.5%) violate it–significantly better than the instantiation with u=s and direct realization (henceforth, DS). On the other hand, the number of utterances with more than one CB almost doubles again (and with respect to the DS instantiation), to 26 (3.9%) from 14 (2.1%).

The overall statistics about pronominalization with the IS instantiation are as follows:

[70]The results in this section were obtained by setting both `realizes` to `indirect` and `uttdef` to `s`.

	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

	MUSEUM	PHARMA	TOTAL
Total number of realizations of CBs:	172	160	332
Total number of CBs realized as R1-pronouns:	86	43	129 (38.9%)
CBs realized as personal pronouns:	86	43	129 (38.9%)
CBs realized as demonstrative pronouns:	4	1	5 (1.5%)
CBs NOT realized as R1-pronouns:	86	117	203 (61.1%)
Total number of R1-pronouns that do not realize CBs:	53	19	72 (33.2%)
Personal pronouns that do not realize CBs:	53	19	72 (33.2%)
Demonstrative pronouns that do not realize CBs:	3	12	15 (65.2%)

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. The complete figures about violations and verifications for the three versions of Rule 1 follow.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	166	198	364 (93.3%)
GJW 95 - utterances that violate:	18	8	26 (6.7%)
Gordon - utterances that satisfy:	60	37	97 (24.9%)
Gordon - utterances that violate:	124	169	293 (75.1%)
GJW 83 - utterances that satisfy:	132	132	264 (67.7%)
GJW 83 - utterances that violate:	52	74	126 (32.3%)

With this instantiation, as well, we would get significantly worse results for Rule 1 if we were to assume that Rule 1 applies to demonstrative pronouns as well, as 75% of demonstrative pronouns do not realize CBs, confirming again the findings, e.g., of (Passonneau 1993).

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%). Complete results are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments :	45	59	104 (15.5%)
Continuations :	27	36	63 (9.4%)
Retain :	52	62	114 (17.0%)
Smooth Shift :	10	25	35 (5.2%)
Rough Shift :	60	40	100 (14.9%)
Zero :	41	57	98 (14.6%)
Null :	36	119	155 (23.2%)

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). The Kibble score is 1.95, much better than with DS (1.4) and IF (1.6).

Treating bridging references as containing null traces Some readers might think that the additional violations of Rule 1 obtained with instantiations IF and IS (such as the one in example (30)) 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.[71] Notice that's what at stake here is not the underlying semantics of bridging references–we agree with this view of their semantics–but whether these implicit anaphors are governed by Rule 1. I.e., the issue is the same raised by relative traces. However, it turns out that treating these implicit anaphoric references as R1-pronouns[72] actually results in *more* violations of Rule 1, even though it is still verified: 25, 5.2%, with IF (instead of 16, 3.4%) and 29, 7.7%, with IS (instead of 26, 6.7%). 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 becomes a potential violation. In (31), for example, the CB of this utterance, *Rocester*, is referred to by a proper name; if we assume that *a few made of bronze*, an (intrasentential) bridging reference to *two shale bracelets*, contains a (null) pronoun, the utterance becomes a violation of Rule 1.

(31) Two shale bracelets were found at Rocester, as well as a few made of bronze

The one positive aspect of this instantiation is that it is the first among those discussed that verifies the version of Rule 1 proposed by (Gordon et al. 1993), both in the IF instantiation (+339, -142, $p \leq .01$ by the sign test) and in the IS one (+247, -128, $p \leq .01$).

Treating the implicit references in bridges as R1-pronouns - hence, as CFs - also has negative effects for Strong C1 and Rule 2, in that it leads to a dramatic increase in the number of utterances with more than one CB (to 100 - 10.3% - with IF, and 94 - 14.1% - with IS), as well as in the number of Rough Shifts, which become 18.6% of all transitions with IF, and 23.6% with IS. These results suggest that these traces are best *not* treated as R1-pronouns; as a result, this is the parameter setting we have used in the remaining experiments.

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).[73] 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[74] 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), allowing entities realized as 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

[71] Such treatments of bridging references have been proposed, e.g., in (Barker 1991; Poesio 1994).

[72] This can be done by setting the parameter `bridges_policy` ('Treatment of Bridges') to `somers`.

[73] 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.).

[74] This can be done by removing `per2` from the value of `cfselect`, the list of NPs not treated as being realized as CFs.

the constraint verified (by the sign test, $p \leq .03$). (The improvement is significant: with 96 former violations being eliminated and only 5 new ones, $p \leq 0.01$.) 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. The overall figures for transitions in the IF instantiation are shown in the following table:

	MUSEUM	PHARMA	TOTAL
Establishments:	75	121	196 (20.2%)
Continuations :	49	79	128 (13.2%)
Retain :	76	60	136 (14.0%)
Smooth Shift :	39	37	76 (7.8%)
Rough Shift :	60	57	117 (12.0%)
Zero :	60	82	142 (14.6%)
Null :	46	131	177 (18.2%)

whereas for the IS instantiation the figures are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	45	52	97 (14.5%)
Continuations :	27	65	92 (13.8%)
Retain :	52	67	119 (17.8%)
Smooth Shift :	10	36	46 (6.9%)
Rough Shift :	60	63	123 (18.4%)
Zero :	41	50	91 (13.6%)
Null :	36	65	101 (15.1%)

There are no significant changes with Rule 2 (Strube and Hahn). Finally, the Kibble coefficient increases for all instantiations: 1.52 for DF (vs. 1.15), 1.83 for DS (vs. 1.4), 1.91 for IF (vs. 1.6), and 2.3 for IS (vs. 1.95).

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 (i.e., if we keep the `prodef` parameter set to `narrow`, as done until now), 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,[75] the

[75]This is done by setting the `prodef` parameter ('Notion of pronoun used') to `two`.

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.(For DF, 154 PRO2s refer to the CB, whereas 146 do not; for IS, 127 PRO2s refer to the CB, whereas 141 do not.) 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 considered so far about which entities to realize 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[76] does not have any positive result at all; on the contrary, it does have a significant (if small) negative impact on Strong C1[77] 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*.[78]

(32) 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.[79]

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, especially with u=s instantiations. 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;

[76]This can be done by adding `predicate` to the value of the `cfselect` parameter–the list of NPs not to be treated as CF realizations.

[77]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.

[78]Part of the problem here is the underlying semantics of copular clauses. Some favor a 'uniform' view in which the postcopular NP is always treated as a predicate. However, in these examples it might be argued that it is the subject NP that expresses the predication; in some other cases, copular clauses can be viewed as stating an equality (e.g., in the infamous *The temperature is ninety*).

[79]The impact on R1 is also negative, but not significantly so.

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.

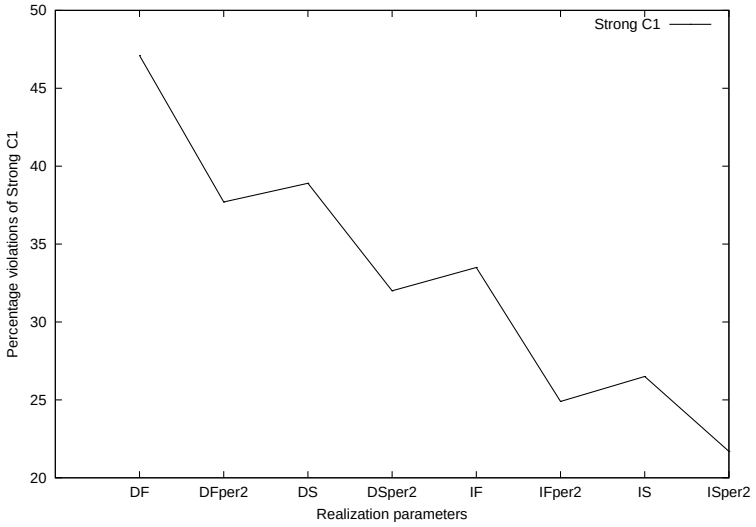

Figure 3: The effect of the realization parameters on the violations of Strong C1.

4.4 Segmentation

As mentioned above, in the main group of experiments reported in this paper we didn't study the effect on the claims made by Centering of different theories about segmentation and the global focus.[80] All we did was to compare different heuristic methods for segmenting a text.[81] In addition, we carried out a second series of experiments using a different corpus to examine the consequences of using a semantic definition of subordination based on rhetorical annotation, rather than the 'syntactic' notion of subordination used here. These experiments are discussed in Section §6.

The result of this comparison is what one would expect: the smaller the segment, the better the results for Strong C1, since having more segments increases the number of utterances at segment boundaries, and these are not counted as violations. So, the number of violations of Strong C1 increases progressively as segment size increases. When sections are treated as segments, the Constraint only holds for the instantiations with indirect realization, or if PRO2s are considered as realizations of CFs.

[80] Such as, say, the differences resulting from using an RST-based segmentation (Hobbs 1979; Reichman 1985; Mann and Thompson 1988), one based on the intentional structure of the discourse (Grosz and Sidner 1986), and one entirely based on entity coherence (Knott et al. 2001); or, among theories based on the assumption that the structure of a discourse is primarily dependent on its intentional structure, between stack-based theories and cache-based ones (Walker 1998). See, however, (Poesio and Di Eugenio 2001).

[81] The parameter `segment` ('Segmentation Heuristic') controls this aspect of the system.

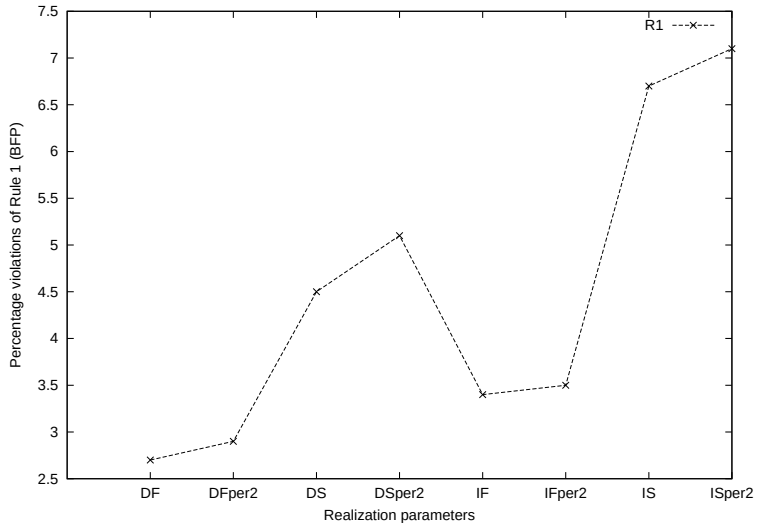

Figure 4: The effect of the realization parameters on the percentage of violations of Rule 1.

When entire texts are treated as single segments, Strong C1 only holds for the IF and IS instantiations– for none of which, however, Rule 2 in the BFP version holds. The validity or otherwise of Rule 2 in the BFP version is not affected by changes in segment granularity, but larger segments lead to worse results both with Strube and Hahn's version (most segment boundaries become expensive transitions) and Kibble's version- e.g., when entire texts are treated as single segments, the Kibble Score goes down to 1.83 for IS (from 2.3) and to 1.6 for IF (from 1.95). R1 is unaffected by the size of the segment, since all that matters is which entity is the CB, and segmentation doesn't affect that.

More specifically, treating every paragraph as a separate segment, instead of using Walker's more relaxed heuristic, doesn't affect the results, as no paragraph in our corpus contains a pronoun referring to an entity introduced in a previous paragraph. Treating each section of a text as a separate segment leads to significantly worse results for DF, and Strong C1 only holds at the .03 level for the instantiation with PRO2s as CFs (+464, -409), not for any other. With DS, Strong C1 only holds for the instantiations with PRO2s as CFs. Strong C1 holds for all instantiations with indirect realization. With no segmentation at all, the results were significantly worse for all instantiations; as a result, Strong C1 is only verified with the IF and IS instantiations.

4.5 Ranking

In this section we examine the effect of changes in the ranking function, controlled by the `ranking` parameter.

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 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 easily 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 et al. 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. The first ranking function we evaluate is GFTHERELIN, obtained by adding linear order to grammatical function.[82] The results with this ranking function are summarized in Table 13.[83]

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 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

[82]The reason for the 'there' in the name 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) for evidence that such NPs are highly ranked.)

[83]To reproduce these results, use the value `gftherelin` for the `ranking` parameter.

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 (33).

(33) (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 full results for the three versions of Rule 1 under the IF instantiation (without predicative NPs and PRO2s) are as follows:

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	239	264	463 (96.3%)
GJW 95 - utterances that violate:	15	3	18 (3.7%)
Gordon - utterances that satisfy:	86	55	141 (29.3%)
Gordon - utterances that violate:	168	172	340 (70.7%)
GJW 83 - utterances that satisfy:	169	151	320 (66.5%)
GJW 83 - utterances that violate:	85	76	161 (33.5%)

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). The overall results for all three versions of Rule 1 under the IS instantiation not including predicative NPs and PRO2s are shown in the following table.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	166	212	378 (92.6%)
GJW 95 - utterances that violate:	21	9	30 (7.4%)
Gordon - utterances that satisfy:	58	43	101 (24.8%)
Gordon - utterances that violate:	129	178	307 (75.2%)
GJW 83 - utterances that satisfy:	129	137	266 (65.2%)
GJW 83 - utterances that violate:	58	84	142 (34.8%)

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%. The complete figures with the DF instantiation are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	91	100	191 (19.7%)
Continuations :	39	36	75 (7.7%)
Retain :	22	17	39 (4%)
Smooth Shift :	18	13	31 (3.2%)
Rough Shift :	13	3	16 (1.6%)
Zero :	67	74	141 (14.5%)
Null :	155	324	479 (49.3%)

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. In comparison with the DS instantiation with simple grammatical function, we find a small increase in CON, and a small decrease in RET and RSH; only this latter is significant. With the IF instantiation, again we have a small increase in CON. but a rather more significant decrease in RSH (from 9.7% to 7.6%) and a small increase in RET. The overall figures for IF are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	74	95	169 (17.4%)
Continuations :	49	44	93 (9.6%)
Retain :	79	56	135 (13.9%)
Smooth Shift :	34	24	58 (6.0%)
Rough Shift :	48	26	74 (7.6%)
Zero :	66	78	144 (14.8%)
Null :	55	244	299 (30.8%)

Finally, with the IS instantiation, we get almost the same results as with IF: 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. As for the Kibble score, it is increased under all instantiations, but by a very small amount: from 1.57 to 1.61 for IF without PRO2s, from 1.93 to 1.97 for IF with PRO2s, from 1.91 to 2.01 for IS without PRO2s, and from 2.28 to 2.38 for IS with them.

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 instan-

tiation 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 et al. 1993; Stevenson et al. 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.[84] 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 function than GFTHERELIN;[85] 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.

Combining Grammatical Function and Linear Order The experiments by Gordon et al. (1993) suggest that CFs in subject and first-mention position have equal ranking. We tried therefore a ranking function in which the first mentioned entity and the subject are equally ranked, then everything else is ranked according to grammatical function; and one in which the first-mentioned entity is always ranked most highly, then the subject, then everything else. With these ranking functions we obtain results comparable to those obtained with simple grammatical function and with GFTHERELIN; which is not terribly surprising, given that we just saw that in our corpus the results with linear order and grammatical function are pretty similar.

We concentrate here on the unambiguous form of this ranking function, in which first-mentioned entities are ranked higher than subjects.[86] Again, no differences were observed (or expected) for Strong C1. Small but not significant differences were observed with R1, and generally in favour of the Gordon *et al.* proposal. The one example which resulted in a violation of Rule 1 with the DF instantiation and ranking=GFTHERELIN, but not with the combined ranking, is the following, in which *Sieber-Fuchs* is pronominalized in (u2).

(34) (u1) For Sieber-Fuchs, old pill packaging, sweet wrappers or photographic film (5), create rich possibilities of colour and texture,

(u2) and she weaves these unlikely materials into bold and exotic jewellery.

[84] 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.

[85] This point is reinforced by a number of results from Gordon and collaborators (e.g., (Gordon et al. 1993, 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.

[86] This setting can be tested using `gordon` as the value of the parameter `ranking`

(Notice that *Sieber-Fuchs* is the CB in this case because of the added order-based disambiguation.)

The results for R1 with the four instantiations are as follows: with DF, +293, -10 (+1); DS, +258, -16 (-2); IF, +464, -17 (+1); IS, +376, -32 (-2).

The results concerning Rule 2 with this instantiation are again pretty similar to those obtained with GFTHERELIN; but, as in the case of pure linear order, every metric is very slightly worse. A few transitions classified as CON become RET, and a few others change from SSH to RSH. The Kibble scores are all very slightly lower: e.g., KS=1.98 for IS (versus KS=2.02 with GFTHERELIN). The percentage of cheap transitions is also lower throughout–e.g., with IS we have Cheap=141, Expensive=303 (vs. 146 and 298).

Information Structure Replacing GFTHERELIN with the ranking function proposed by Strube and Hahn (1999), henceforth, STRUBE-HAHN[87] 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.[88] 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 (20), illustrating the kind of situations in which a partial ranking may result in two CBs. We repeat that sentence in (35), including the preceding sentence.

(35) (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.

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, (36) is an example in which GFTHERELIN ranking results in a violation of Rule 1, while STRUBE-HAHN ranking doesn't.

(36) (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 (35) and (36) are almost stereotypical instances

[87] As discussed earlier, this function ranks HEARER-OLD entities more highly than INFERRABLES, and these higher than HEARER-OLD entities. This function can be tested using the `strube` setting of the `ranking` parameter.

[88] We only discuss the results with the version of Rule 1 proposed by Grosz et al. (1995).

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). With the IS instantiations, 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.[89] The transition percentages with IS and STRUBE-HAHN ranking are in Table 15.

	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, and given the type of comparison we are running, 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 GFTHERELIN. Adopting STRUBE-HAHN ranking does result in a greater percentage of utterances being classified into one of the 'continuous' classes proposed in Grosz et al. (1995); Brennan et al. (1987), 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.

4.6 Other definitions of CB

Gordon et al, 1993 The definition of CB proposed by Gordon et al. (1993) is perhaps the one that makes the trade-off between Strong C1 and R1 most evident. With this definition we find a dramatic increase in the number of utterances without CB; but also a dramatic reduction in the number of violations to R1.

With the DF instantiation, using (Gordon et al. 1993) definition of CB and the ranking function proposed in that paper, 59% of utterances have no CB, as opposed to 46.3% with the same instantiation, GFTHERELIN, and the definition of CB provided by Constraint 3. However, the number of violations of R1 goes down from 11 to 2, also a significant improvement. Most of these are simply

[89] 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.

utterances that do not have a CB according to the definition of Gordon *et al.*; however, in three cases we see a genuine improvement. One of these cases is (34), already seen above, where Sieber-Fuchs becomes the CB of (u2).

(34) (u1) For Sieber-Fuchs , old pill packaging , sweet wrappers or photographic film (5) , create rich possibilities of colour and texture ,

(u2) and she weaves these unlikely materials into bold and exotic jewellery .

With this instantiation, Gordon *et al.*'s own version of Rule 1 (always pronominalize the CB) is verified (58.3% of utterances verify it); but the example below illustrates the fact that even this new definition doesn't always result in pronouns referring to the CB. *This new title* is the CB in (u311), but it's not pronominalized.

(37) (u307) In the same year, he achieved the title of cabinetmaker and sculptor to Louis XIV, King of France. (u311) This new title allowed him to produce furniture as well as works in gilt bronze such as chandeliers, wall lights, and mounts.

This last example is a very clear illustration of the phenomenon observed, e.g., by Brennan (1995): in some cases, a discourse entity has to be moved into a more salient position before it can be pronominalized; simply being the only entity from the previous utterance mentioned in the current one doesn't appear to be sufficient.

The reduction in the number of violations of Rule 1 (GJW 95) is even greater for the u=s instantiations. With DS we find only 3 violations, 2.6%, 12 fewer than with the configuration using the definition of CB from Constraint 3, but Gordon *et al.*'s version of Rule 1 is not verified. Even larger reductions in the number of violations of R1 are found with the IS instantiation: just 3 (2.5%), as opposed to 31 (7.6%) with the 'classic' definition of CB in Constraint 3 and GFTHERELIN. However, with this definition of CB only very few utterances –119, 17.8%–have one, and therefore may verify or violate Rule 1.

Examining the results concerning Rule 2, another consequence of adopting Gordon *et al.*'s definition of CB comes to light: a virtual elimination of all types of transitions apart from continuations and establishments. With the IF instantiation, for instance, 13.6% of utterances (132) are Establishments, 3% are Continuations, and we find only 2 RET, 17 SSH, and 2 RSH. 72% of utterances are NULL.

Passonneau We tested two versions of Passonneau's proposal: one in which the CB is only established if we have strong parallelism between the two pronouns - i.e., they have the exact same grammatical function–and one in which only two types of position are considered: 'SUBJECT' and 'OTHER'. The results with this definition of CB can be summarized as follows: very few utterances end up having a CB (more precisely, a Local Center); but once the CB is established, nothing else gets pronominalized in its stead. So, for example, with the DF instantiation and GFTHERELIN ranking, only 22 utterances (2.3%) have a CB, but there are no violations of Rule 1, irrespective of which version of R1 one adopts. Yet, the usefulness of the notion of 'center' to predict pronominalization is completely lost: for while it is true that 25 out of 27 realizations of a CB are done using personal pronouns, 91.1% of personal pronouns are not realizations of CB; so a separate story would have to be told to account for the cases (the great majority) of pronouns not referring to the Local Center. An example in which the CB / Local Center does get established (in (u64)) is the following:

(38) (u62) The fleur-de-lis on the top two drawers indicate that the cabinet was made for Louis XIV.

(u63) As it does not appear in the inventories of his possessions,

(u64) it may have served as a royal gift.

The same results are obtained when we vary the utterance and realization parameters: with no instantiation we find more than 3% utterances with a CB, or any violation of Rule 1, or that more than 15% of personal pronouns are CB realizations.

Looking at transitions, we find with all instantiations that at least 90% of utterances are NULLs, and of the utterances that do have a CB, most are establishments, and the rest continuations, except for a single retain with IF, DF, and DS. With DF, for example, we find 933 NULL (96%), 21 establishments (2.2%), 1 continuation, and 17 zero - no shifts, and no retains.

As it turns out, the results are slightly better if we allow for a looser notion of parallelism, but not dramatically so. We still don't have any violations of R1 under any of the definitions we are considering; and a few more utterance have a CB, but the difference is not significant (e.g., 23 instead of 20 for DF, and 22 instead of 18 for DS). Both with DF and DS around 90% of pronouns still do not refer to the Local Center.

5 LINGUISTIC CLAIMS BASED ON CENTERING

The notion of 'topic' plays an important role in linguistic theories concerning discourse effects on syntax. These include, e.g., theories of the factors affecting the choice of NP form (Givon 1983; Ariel 1990; Gundel et al. 1993), especially concerning the use of empty subjects or objects in languages that allow them (Kameyama 1985; Walker et al. 1994; Di Eugenio 1998; Prince 1998); or theorizing about languages in which 'topics' occupy fixed positions or topichood licences certain types of movements, such as scrambling (Vallduvi 1990; Rambow 1993; Portner and Yabushita 1998). One of the reasons for the interest in Centering among linguists is the possibility that the theory may make the elusive notion of 'topic' more precise, and such claims easier to verify. Conversely, these claims may serve as a different type of evaluation of the theory, thus serving as a different way of identifying the 'best' parameter instantiation–the 'best' instantiation being the one which makes more of such claims verified. In this section, we use the results just presented concerning the 'best' ways of setting the parameters of Centering to examine these linguistic claims.

5.1 Demonstratives

A number of studies suggest that *this*-Noun Phrases–demonstrative pronouns *this* and *these*, and full NPs with *this* as a determiner–are primarily used to refer to entities that are somehow 'salient', but without being the 'focus' or 'topic' of the discourse. Examples of entities that are felicitously realized by means of THIS-NPs are entities in the visual situation, or 'deixis', such as 'the room' in (39) (Kaplan 1979; Jarvella and Klein 1982; André et al. 1999); abstract objects such as propositions, facts, or types, implicitly introduced in the discourse without being explicitly mentioned, as in (40) (Asher 1993; Webber 1991); and entities mentioned in a discourse, but not in most salient position, such as 'the area' in (41) (Linde 1979; Gundel et al. 1993; Passonneau 1993).

(39) A [inside a room, looking around]: This room is incredibly dirty.

(40) For example, binocular stereo fusion is known to take place in a specific area of the cortex near the back of the head. Patients with damage to this area of the cortex have visual handicaps but they show no obvious impairment in their ability to think. This suggests that stereo fusion is not necessary for thought. (Webber 1991)

(41) a. In spite of his French name, **Martin Carlin** was born in Germany and emigrated to Paris to become an ébéniste.

b. **He** settled there with other German and Flemish craftsmen and took employment in the workshop of Jean-Fran cois Oeben, whose sister **he** married.

c. Inventories made after **Carlin**'s death show that **the ébéniste** and his wife lived modestly in a five-room apartment in THE FAUBOURG SAINT-ANTOINE, an unfashionable quarter of Paris, with simple furniture, a few pastel portraits, and a black lacquer clock.

d. Few of **Carlin**'s wealthy clientele would have cared to venture into THIS AREA

Gundel et al. (1993) proposed that the NP form chosen to realize discourse entities results from the interaction of two factors: the speaker's assumptions about the status of such entities in the addressee's 'mental status', and Grice's Maxim of Quantity (Grice 1975) requiring speakers to be as informative as consistent with their knowledge, but no more informative. Gundel *et al.* propose that there are six possible 'Givenness statuses', organized in a 'Givenness Hierarchy' reflecting increasing 'mental salience'. The statuses relevant to the distribution of THIS-NPs are the highest two: ACTIVATED and IN FOCUS. In order for a THIS-NP to be felicitous, the discourse entity it realizes is required to be at least ACTIVATED: Gundel *et al.* define 'activated' as 'being in short-term memory'. On the other hand, if an entity was 'in focus', we would expect the speaker to be as informative as possible and therefore use a pronoun, the type of NP typically used to realize such entities, unless other factors made the use of a pronoun infelicitous.

Gundel (1998) already proposed that notions from Centering can be used to specify more precisely which entities may be 'in focus', although she also calls for a modification of the theory, arguing that more than one entity may be 'in focus' (we return to this topic in the Discussion). Poesio and Nygren-Modjeska (2002) devised a reliable scheme to annotate 'activated' entities, and identified three 'natural' ways of formalizing the notion of entity 'in focus' used by Gundel *et al.* in terms of the conceptual vocabolary of Centering. An entity may be 'in focus' if it is:

1. CB(U_i), the CB of the present utterance.
2. CB(U_{i-1}), the CB of the *previous* utterance.
3. CP(U_{i-1}), the most highly-ranked entity of the *previous* utterance.

Poesio and Modjeska analyzed every THIS-NP in the corpus used in this study, using the script used for this study to compute utterances and their CB and CP according to four instantiations: IF and IS, with both GFTHERELIN and STRUBE-HAHN ranking. Their analysis revealed that virtually all THIS-NPs in the corpus were activated, and that between 90 and 93% of THIS-NPs (depending on the instantiation) were used to refer to entities other than CB(U_{i-1}); between 75-80% to refer to entities other than CP(U_{i-1}); and between 61-65% to refer to entities other than CB(U). They concluded that the distribution of THIS-NPs in the corpus used for both this study and their is best characterized by what they called the THIS-NP hypothesis:

THIS-NP **Hypothesis** : THIS-NPs are used to refer to entities which are ACTIVATED (in the sense specified in the paper). These entities should not be CB(U_{i-1}).

5.2 Type of Transition vs Form of Subject

Kameyama (1986); Di Eugenio (1998); Turan (1998) argued that in languages with both a 'weak' and a 'strong' pronominal form, the form of the subject of an utterance is affected by the type of transition (CON, RET, etc.) that that utterance realizes. Typically, it was argued, weak pronominal forms are preferred with center continuations, whereas strong pronominal forms are preferred for center shifts and center retains. In the case of English, Passonneau and others found a similar correlation between CON and personal pronouns, whereas other transitions correlated more with demonstrative pronouns. In this section we discuss our results concerning these correlations with the 'best' instantiations identified above. However, because of the low frequency of some events,[90] our results should be considered as preliminary.

IS, GFTHERELIN: The full contingency table for the instantiation IS, with ranking function GFTHERELIN, is as follows:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	10	1	77
CON	21	3	30
RET	2	3	79
SSH	9	0	13
RSH	3	1	60
ZERO	0	3	47
NULL	2	1	53
TOTALS:	47	12	359

This contingency table cannot be used for a χ^2 test, because of the low or zero counts in some of the cells; we need to collapse some of the distinctions between transitions. One possibility is to collapse the SSH and RSH cells; another is to collapse demonstrative NPs with full NPs. We then obtain the following contingency table:

	PRONOUN	FULL NP	TOTAL
EST	10	78	88
CON	21	33	54
RET	2	82	84
RSH-SSH	12	74	86
ZERO	0	50	50
NULL	2	54	56
TOTAL	47	371	418

This table still contains cells with values under 5, however, which tend to increase the χ^2 value, so more drastic merges are required. One possibility is to merge CON and RET (both of which continue the same CB) and RSH-SSH with ZERO (both of which lead to a change in CB). However, the resulting contingency table still contains low values in the ZERO row. Another way of eliminating the low counts is to simply drop RET and NULL, and to merge ZERO and SSH-RSH:

[90] Not all cells of our contingency table contain at least 5 elements, which increases the chances of a Type 1 error (Woods et al. 1986), so we consider various ways of reducing its dimensionality.

	PRONOUN	FULL NP	TOTAL
EST	10	78	88
CON	21	33	54
RSH-SSH-ZERO	12	124	136
TOTAL	43	235	278

This new table doesn't have low count cells, and the distribution is highly significant: $\chi^2 = 27.1, p \leq 0.000$. Given that (Walker et al. 1994) argue that EST and CON are the same transition, one might also think of collapsing together the EST and CON rows–although this merge does not look very promising, since different types of NPs may be used to turn a discourse entity into the CB and to continue the current CB. The resulting contingency table is as follows:

	PRONOUN	FULL NP	TOTAL
EST / CON	31	111	142
RSH-SSH-ZERO	12	124	136
TOTAL	43	235	278

This distribution is significant: $\chi^2 = 8.99, p \leq .003$. However, treating EST as a type of SHIFT results in a much higher χ^2 value. The resulting distribution is shown in the following contingency table:

	PRONOUN	FULL NP	TOTAL
CON	21	33	54
RSH-SSH-ZERO-EST	22	202	224
TOTAL	43	235	278

This distribution is highly significant: 1df, $\chi^2 = 28.1, p = .000$. This χ^2 value is much higher not only than the value obtained by gtreating EST as a type of continuation, but also than the value obtained by merging CON and RET, as in the contingency table below, which is not significant at the .05 level ($\chi^2 = 3.68$).

	PRONOUN	FULL NP	TOTAL
CON-RET	23	115	138
RSH-SSH-ZERO-EST	22	202	224
TOTAL	45	317	362

IF, GFTHERELIN**:** With the IF instantiation and GFTHERELIN ranking, we get the following contingency table:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	25	1	94
CON	36	2	32
RET	0	1	98
SSH	17	0	9
RSH	4	4	56
ZERO	0	3	77
NULL	8	4	82
TOTALS:	90	15	448

Again, collapsing RSH and SSH, and the two columns DEM and FULL, is not enough to completely eliminate the low counts. However, after eliminating RET as well, merging ZERO with the shifts, and eliminating NULLs, we get a contingency table with sufficient counts in all cells:

	PRONOUN	FULL NP	TOTAL
EST	25	95	120
CON	36	34	70
RSH-SSH-ZERO	21	149	170
TOTAL	82	278	360

For this table, $\chi^2 = 43.4$, which with 2df is highly significant.

IS, STRUBE-HAHN: With the ranking function proposed by Strube and Hahn, we get the following contingency table:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	10	1	77
CON	22	5	50
RET	2	1	62
SSH	10	0	25
RSH	1	1	45
ZERO	0	3	47
NULL	2	1	53
TOTALS:	47	12	359

Again, in order to ensure enough elements in all cells, it is necessary to collapse the columns for demonstratives and full NPs, merge the rows for RSH, SSH, and ZERO, and drop the rows for RET and NULL. The results are shown in the following contingency table:

	PERS PRONOUN	FULL NP	TOTAL
EST	10	78	88
CON	22	55	77
SSH-RSH-ZERO	11	121	132
TOTALS:	43	254	297

This distribution is significant ($\chi^2 = 17.1$).

IF, STRUBE-HAHN: The complete contingency table for this instantiation is as follows:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	25	1	94
CON	35	2	55
RET	0	1	82
SSH	19	0	21
RSH	3	4	37
ZERO	0	3	77
NULL	8	4	82
TOTALS:	90	15	448

Collapsing as above, we get the following distribution, which is highly significant ($\chi^2 = 21.1$).

	PRONOUN	FULL NP	TOTAL
EST	25	95	120
CON	35	57	92
RSH-SSH-ZERO	22	142	164
TOTAL	82	294	346

Summary A dependency between the three-way distinction between types of transition (EST / CON / RSH-SSH-ZERO) and the form of subject NP (pronoun or full NP, counting demonstrative pronouns among the full NPs) was observed. The dependency was shown to be significant by a χ^2 test for all four instantiations shown to be 'best' by the analyses in Section §4. Our results indicate that continuations are best kept separate from retains. They also suggest that for the purpose of predicting the form of the subject, it's not a good idea to view establishments as a type of continuation, as suggested in (Walker et al. 1994); establishments appear to behave more like shifts.

We should note however that the correlation suggested by the χ^2 test is only a tendency, so our results don't necessarily translate in good algorithms for deciding the form of NP to be used in subject position depending on the transition; this point is illustrated more concretely below when discussing the correlation between transitions and segment boundaries.

5.3 The correlation between transitions and segment boundaries

Passonneau (1998) and Walker (1998) studied whether transitions predict segment boundaries, i.e., whether establishments and shifts occur more at segment boundaries, and continuations prevail within a segment. These studies didn't find much of a correlation, but only one instantiation of the theory was considered; we tried therefore to see if we could get a better result using the 'best' instantiations identified above. Again, readers should keep in mind that our analysis can only be viewed as indicative, the more so given that our corpus wasn't properly annotated for segments.

IF, GFTHERELIN**:** The relation between transitions and boundaries with this instantiation is shown in the contingency table below:

	NOT BOUNDARY	BOUNDARY	TOTALS:
EST	115	54	169
CON	75	18	93
RET	119	16	135
SSH	54	4	58
RSH	65	9	74
ZERO	98	46	144
NULL	221	78	299
TOTALS:	747	225	972

The results of the χ^2 test for this table are highly significant: with 6df, $\chi^2 = 39.1, p \leq .001$. It is nevertheless clear from the table that the correlation between CB continuations and segment continuations is imperfect at best: the percentage of continuations that are boundaries (about 19%) is not that much lower than the overall percentage of boundaries (about 24%). But perhaps the most interesting

aspect of this table is that it gives further support to the hypothesis that establishments behave differently from continuations, formulated above while discussing the correlation between transition type and subject type. Whereas slightly less that 1/5 of continuations are segment boundaries, almost 1/3 of EST are, a higher percentage than that for utterances overall (1/4). In fact, EST are more frequently boundaries that the shifts or ZERO; the only other transition that correlates as highly with segment boundaries is NULL (1/3 of NULL are boundaries, as well). This suggests collapsing the categories as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	194	34	228
SSH+RSH+ZERO+EST+NULL	553	191	744
TOTALS:	747	225	972

This table makes the correlations clearer: boundaries are 23% of the total number of transitions, but 15% of the total number of CON+RET transitions, and 25.5% of the remaining transitions. The χ^2 for this table is $\chi^2 = 11.1, p \leq 0.001$.

IS, GFTHERELIN**:** The results with this instantiation are similar to those just seen with IF and are summarized by the following table:

	NOT BOUNDARY	BOUNDARY	TOTALS
EST	58	49	107
CON	49	23	72
RET	95	26	121
SSH	25	7	32
RSH	60	16	76
ZERO	56	45	101
NULL	101	59	160
TOTALS:	444	225	669

Again, we have the interesting (although not altogether surprising) result that EST are much more frequently boundaries than CON, and NULL are more frequent boundaries than the two types of shift. And again, the distribution is already significant for the table just seen (with 6df, $\chi^2 = 28.7, p = 0$). A collapsed table with only two classes, CON+RET vs. everything else, again makes the correlations more obvious: boundaries are 25.4% of CON+RET, but 37% of the rest. (The χ^2 value for this table is 8.26, lower than with IF.)

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	144	49	193
EST+SSH+RSH+ZERO+NULL	300	176	476
TOTALS:	444	225	669

IF, STRUBE-HAHN: The results with this instantiation are not very different from those obtained with GFTHERELIN. The overall distribution is as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
EST	115	54	169
CON	98	19	117
RET	104	15	119
SSH	53	4	57
RSH	58	9	67
ZERO	98	46	144
NULL	221	78	299
TOTALS:	747	225	972

This distribution is highly unlikely to be due to chance, just like the one with GFTHERELIN: $\chi^2 = 37.5, p \leq 0.001$. The 'collapsed' distribution is as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	202	34	236
EST+SSH+RSH+ZERO+NULL	545	191	736
TOTALS:	747	225	972

Again, given this contingency table it is highly unlikely that the two variables are independent, $\chi^2 = 13.4, p \leq 0.001$. Both of these χ^2 values are very similar to those obtained with GFTHERELIN.

IS, STRUBE-HAHN: Again, the results are similar to those obtained with the IS parameter setting and GFTHERELIN, except that the values of χ^2, while still significant, are lower. The full contingency table is as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
EST	58	49	107
CON	66	28	94
RET	81	22	103
SSH	35	9	44
RSH	47	13	60
ZERO	56	45	101
NULL	101	59	160
TOTALS:	444	225	669

This table has $\chi^2 = 28.01, p \leq 0.001$. Collapsing rows as above gives us the following contingency table:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	147	50	197
EST+SSH+RSH+ZERO+NULL	297	175	472
TOTALS:	444	225	669

The χ^2 value for this table is 8.52.

Summary The tests above indicate that it is very unlikely that the variables TRANSITION and BOUNDARY are independent. This does not mean, however, that transitions are a very good cue for detecting segment boundaries. This can be seen by using the technique proposed by Passonneau in her study (1998). Passonneau measures the usefulness of transitions as cues for segmentation in terms of precision, recall, and ERROR RATE. She uses two classification systems for transitions: the one due to (Grosz et al. 1995) that divides them into CON, RET and SHIFT, and one proposed in (Kameyama et al. 1993) that classifies them into RET1 (= CON+RET), EST (our EST), and NULL (our NULL). Defining error rate E=(CON at boundary + SHIFT at nonboundary)/total, Passonneau gets the following values: for SHIFT as predictor of boundary, R=.78, P=.25, E=.41; for NULL, R=.86, P=.26, E=.40.

Using the 'best' instantiations and the collapsed classes discussed above (CON+RET, EST+ZERO +SSH+RSH+NULL) we get results comparable to Passonneau's. With IF settings, and using the class EST+SSH+RSH +ZERO+NULL to predict boundaries, we get R= 186/225 = .83, P=186/689=.27, and E=39+502/972=.56 with GFTHERELIN ranking ; R = .85, P = .26, E=34+545/972=.59 with STRUBE-HAHN ranking. Using CON+RET to predict nonboundary, the results are R=.33, P=.86, and E as above with GFTHERELIN ranking; with STRUBE-HAHN ranking, R=.27, P=.86, E again as above. With IS, the results are slightly worse, as might be expected given that already the χ^2 values are lower. Using EST+SSH+RSH +ZERO+NULL to predict boundary, we have R=.78, P=.37 with both ranking functions; using CON+RET to predict non-boundary, we get R=.33, P=.75 again with both ranking functions. In other words, even by using the 'best' instantiations and by collapsing transitions in the best way (which is slightly different from Passonneau – in particular, because EST is joined with SHIFT) we get more or less the results that Passonneau gets, and the predictive power of transitions is not very high.

5.4 Long Distance Pronouns

Hitzeman and Poesio (1998) claimed that it is not sufficient for an antecedent to be available on the stack for the use of a long distance pronoun (a pronoun whose antecedent is not in the previous utterance) to be licensed; it is necessary for the entity to have been a CB. We tested this claim using our data and the best instantiations.[91]

The first, perhaps obvious, finding is that the importance of this issue greatly depends on the definition of utterance. Hitzeman and Poesio assumed that each finite clause was a separate utterance, as suggested by Kameyama; if we adopt this definition, then 17 pronouns out of 217 are long distance, which is the same percentage (8%) found in the corpus used by Hitzeman and Poesio.[92] If we identify utterances with sentences, however, we only get 5 long-distance pronouns.[93]

Hitzeman and Poesio's claim is verified in our corpus as well, both with the IF instantiation and the IS instantiation. With IF, 13 long distance pronouns out of 17 had been CBs and 4 had not, $p \leq .02$ with GFTHERELIN ranking; with STRUBE-HAHN, 14 and 3, respectively. With IS, we find +3, -2 with GFTHERELIN ranking and STRUBE-HAHN ranking, but there is not enough data for a significance test. An even better result, however, was found by weakening the licensing condition to having been a CP rather than a CB: in this case, with IF we have +17, 0, $p \leq .01$ by the sign test with GFTHERELIN ranking, and +16, -1 with STRUBE-HAHN ranking. With IS, the results are +4, -1 with GFTHERELIN

[91]There is no overlap between the texts used in this study and the texts used for the Hitzeman / Poesio study, which were spoken dialogues.

[92]Overall, with this definition of utterance, 1158 anaphoric expressions have their antecedents in the current or previous utterance, and 455 at a distance 2-6; none if farther away.

[93]With this definition, 1242 have their antecedent in the same or previous utterance; for 385 is further away.

ranking, and +5, 0 with STRUBE-HAHN.

6 CENTERING THEORY AND RHETORICAL STRUCTURE

The experiments reported above couldn't study the impact of two factors:

- discourse segmentation: we only did a basic segmentation of the texts based on layout;
- subordination: the syntactic annotation of clauses used for the annotation could be used to classify the *because* clause as subordinate in (42), but not in (43), where the same underlying semantic subordination is not syntactically realized:

  (42) John fell *because Max pushed him*. He was drunk as usual.

  (43) John fell. *Max pushed him*. He was drunk as usual.

In subsequent work (Poesio and Di Eugenio 2001; Poesio et al. 2004), we addressed these limitations using a corpus previously annotated according to RELATIONAL DISCOURSE ANALYSIS (RDA) (Moore and Pollack 1992; Moser and Moore 1996b), a theory of discourse structure that synthesizes ideas from Grosz and Sidner's theory (Grosz and Sidner 1986) with ideas from RHETORICAL STRUCTURES THEORY (Mann and Thompson 1988). This corpus was further annotated for anaphoric information and other properties of noun phrases according to the scheme used for the rest of the corpus. In this section we briefly discuss these experiments. (At the moment, these experiments cannot be replicated from the Web site.)

6.1 Relational Discourse Analysis (RDA)

Relational Discourse Analysis (RDA) (Moore and Pollack 1992; Moser and Moore 1996b) owes to Grosz and Sidner the idea that discourse is hierarchically structured, and that discourse structure is determined by intentional structure; each RDA-segment originates with an intention of the speaker. But in RDA segments have additional internal structure: each segment consists of one CORE, i.e., that element that most directly expresses the speaker's intention, and any number of CONTRIBUTORS, the remaining constituents in the segment, each of which plays a role in serving the purpose expressed by the core. The notions of core and contributor derive of course from the notions of nucleus and satellite in Rhetorical Structure Theory (RST) (Mann and Thompson 1988), which claims that in each "segment" (text span, for RST) one component should be identified as the 'main' one, and the others as secondary. However, in RST there is a distinction between nucleus and satellite for (almost) all RST relations, whereas in RDA a core and contributors are only identified if a segment purpose has been recognized.

In RDA, segment constituents may in turn be other embedded segments, or simpler functional elements: these elements may be either basic UNITS, which are descriptions of domain actions and states, or relational CLUSTERS. Clusters are spans that only involve constituents linked by informational relations.

Unlike G&S's theory and like RST, RDA is based on a fixed number of relations; in particular, RDA assumes four intentional relations – **convince**, **enable**, **concede**, **joint**–and a larger set of informational relations; this latter set is expected to be domain dependent. In the Sherlock corpus, 23 informational relations are used, of which 13 pertain to causality (they express relations between two actions, or between actions and their conditions or effects) (Moser et al. 1996).

1.1 Before troubleshooting inside the test station,
1.2 it is always best to eliminate both the UUT and TP.
2.1 Since the test package is moved frequently,
2.2 it is prone to damage.
3.1 Also, testing the test package is much easier and faster
3.2 than opening up test station drawers.

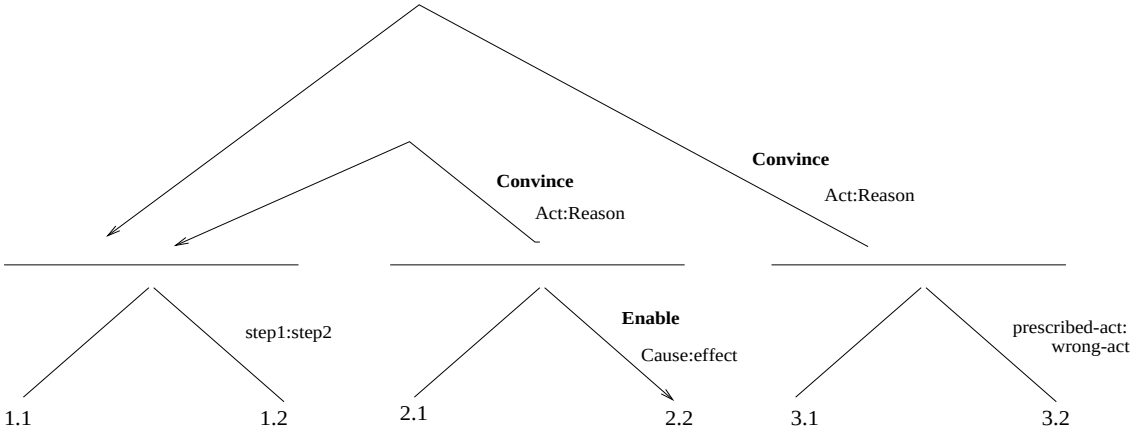

Figure 5: A tutorial excerpt and its RDA analysis

Figure 5 shows a small excerpt from one of the dialogues in the Sherlock corpus, and its corresponding RDA analysis. The text is broken into clauses (UUT is "Unit under test", TP is "test package"). The analysis shows the text to be analyzed as an intentional segment whose core spans 1.1 and 1.2. This segment has two contributors, spanning 2.1 and 2.2, and 3.1 and 3.2 respectively. Graphically, the core is at the end of the arrow whose origin is the contributor; moreover, the link is marked by two relations, intentional (in bold), and informational. In this specific case, the two contributors carry the same intentional and informational relations to the core, but this doesn't need to be the case. The core and the two contributors are further analyzed. The core and the second contributor are analyzed as informational clusters, whereas the first contributor is recognized as having its own intentional structure. Clusters are marked by one informational relation, but not by intentional relations.

6.2 The Sherlock Corpus

The corpus we used for this study is a collection of tutorial dialogues between a student and a tutor, collected within the Sherlock project (Lesgold et al. 1992). The corpus includes seventeen dialogues between individual students and one of 3 expert human tutors, for a total of 313 turns (about 18 turns per dialogue), and 1333 clauses. The student solves an electronic troubleshooting problem interacting with the Sherlock system; then, Sherlock replays the student's solution step by step, schematically criticising each step. As Sherlock replays each step, the students can ask the human tutors for explanations. The student and tutor communicate in written form. Because most of the discourses are explanations, we expected 'relation-based' coherence to play an important role in this corpus.

The Sherlock corpus was previously annotated using RDA to study cue phrases generation (Di Eugenio et al. 1997). The research group which proposed RDA discusses the following reliability results (Moser and Moore 1996a). 25% of the corpus was doubly coded, and the κ coefficient of agreement

was computed on segmentation in a stepwise fashion. First, κ was computed on agreement at the highest level of segmentation. After κ was computed at level 1, the coders resolved their disagreements, thus determining an agreed upon analysis at level 1. The coders then independently proceeded to determine the subsegments at level 2, and so on. The deepest level of segmentation was level 5; the κ values were .90, .86, .83, 1, and 1 respectively (from level 1 to 5).

6.3 Annotation Methodology

We annotated about half of the Sherlock corpus for anaphoric information, using a much simplified version of the annotation scheme used in the previous experiments. More specifically, we marked each NP in the corpus, specified its NP type (proper name, pronoun, the-np, indefinite NP, etc) and grammatical function (subject, object, etc.); and then we marked all 'direct' anaphors between these NPs. We annotated a total of 1549 NPs, 507 of which were anaphoric; 336 NPs were pronouns, of which 48 were third-person. A crucial difference between this study and the ones discussed previously is that we did not annotate 'bridging' information, because without the original drawings of the circuits it was very difficult to determine with certainty which objects were parts of other objects.

6.4 Computing Violations

The annotation thus produced was used to automatically compute utterances, and then to compute the CFs and the CB (if any) of each utterance on the basis of the anaphoric information and according to the notion of ranking specified. (We only considered ranking based on grammatical function.) This information was then used to find violations of Strong C1, Rule 1, and Rule 2.

The main issue we had to consider in this work was how to use rhetorical information to characterize utterances and previous utterances; all previous studies relied on purely syntactic definitions.

- As far as segmentation is concerned, we counted as a segment every RDA-segment, i.e., every span of text for which an intentional 'core' had been recognized. This way of computing segments is fairly generous (i.e., it might result in way too many segments), so should give us a lower bound on the number of violations of Constr. 1.
- We treated each basic unit of the RDA annotation (actions and states, as well as 'matrices' - i.e., verbs with a clausal complement) as a distinct utterance. (Note that these are not all finite clauses.) In total, 784 utterances.
- In clusters (blocks of utterances connected only by informational relations), we used the immediately preceding unit as previous utterance. E.g., in Figure 5, 1.1 was counted as previous utterance for 1.2, and 3.1 as previous utterance for 3.2.
- In segments, we considered two possible choices of previous utterance, on the basis of the suggestions of Kameyama and Suri and McCoy. A unit like 3.1 in Figure 5 could have as previous utterance either (the last unit of) the immediately preceding constituent (i.e., 2.2), or (the last unit of) the dominating element, the core (1.2).

Notice that because the corpus does not contain subordination information in the case of informational relations, we could only explore a subset of all cases of semantic subordination.

6.5 Main Results

In this study we were only really concerned with one parameter, the choice of the previous utterance; but we could also look at whether an improved form of text segmentation changed the results concerning Constraint 1 discussed in the rest of the report. The metric we used to evaluate a particular parameter instantiation was the number of violations of the constraints. The results concerning Strong C1 are summarized in the following table:

	CB	**Segment Initial**	**NO CB**	**Total Number**
Sherlock	76	247	461	784

We found no violations of the versions of Rule 1 proposed in (Grosz et al. 1995) and (Grosz et al. 1983). Of the 48 pronouns, 29 were CBs, 19 weren't; of these 19, 4 were references to actions, 8 were long-distance, 6 intrasentential. Most CBs (47) were not pronominalized. Finally, we evaluated the versions of Rule 2 from (Grosz et al. 1995) and from (Brennan et al. 1987). The figures concerning single transitions are as follows (where we have classified as **Zero** each transition from an utterance with a CB to one without, and as **Null** each transition between two utterances none of which had a CB):

Establishment	61
Continuation	5
Retain	5
Smooth Shift	4
Rough Shift	1
Zero	39
Null	669

There are no sequences of continuations, rough shifts, and of retain followed by any shift; 5 establishment / continuation sequences; and 491 sequences of null transitions.

6.6 Discussion

This experiment confirms one of the findings of the experiments relying on syntactic embedding only: even when using a more accurate annotation for segmentation, it is still the case that with direct realization most utterances have no CB–in only 76 cases (10% of the total) an entity introduced in one utterance is mentioned again in the next utterance. What does change is the number of segment boundaries, much higher than in the experiments discussed in Section §4. The fact that only 10% of utterances have a CB is in part due to the fact that we did not annotate for bridging references, but also to the fact that in this domain relational coherence plays a more important role than it did in the other two domains (see Discussion). For example, utterances a. and b. below do not refer to the same objects (if perhaps very indirectly), but coherence is nevertheless achieved because the first one expresses information that is necessary to support the second. The same is true of c. and b.

(44) a. You know that one of the measurement paths is bad.
b. Showing the UUT, TP and measurement section as unknown is correct
c. because when you get your fail you know that something is wrong.

Our second main result is that using rhetorical units to define utterances, and semantic subordination instead of syntactic subordination to define 'previous utterance', also does not seem to change the result that the two notions of 'previous utterance' proposed in the literature are not significantly different. In fact, we found that blurring the distinction between finite and non-finite clauses is probably not a good idea. However, we could only test a subset of the possible cases of subordination with the present corpus.

7 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).

7.1 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.[94] 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 2003)– 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. (We return to Strong vs. Weak Constraint 1 and entity coherence below.) 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. (We return to this finding as well.)

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

[94] Except for the version of the Rule that one can extract from (Gordon et al. 1993), which however was only meant to apply in a limited range of situations.

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.[95] 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–we saw while discussing the correlation between transitions and the form of NP why this might not be a good idea); indeed, in the IS instantiation with GFTHERELIN 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. (We will return on the claims below.)

The relative importance of parameters Of the parameters of Centering, ranking is the one that has been most extensively discussed. An interesting aspect of our results is that because Strong C1 is the claim with the greater number of violations, the most effective way to obtain a instantiations verifying all claims is to change the setting of parameters that have not been as widely discussed, such as whether utterances should be identified with sentences or finite clauses, and those controlling realization. Another interesting result is that changing the ranking function does not even have an impact on Rule 1. Using linear order instead of grammatical function, or a combination of the two, did not significantly affect the results for either Rule 1 or Rule 2, due to the strong positive correlation between first mention in a sentence and subjecthood. Adopting the ranking function proposed by Strube and Hahn does have an effect, a better distribution of the four 'classical' transitions, but the result–making Rule 2 (BFP version) verified at the .01 instead of the .05 level–is comparatively minor.

Another parameter which has originated some discussion in the literature, the definition of previous utterance (Suri and McCoy 1994; Cooreman and Sanford 1996; Kameyama 1998; Strube 1998; Miltsakaki 1999; Pearson et al. 2000) was also found to have limited impact. Adopting a 'Suri-like' notion of which utterance should be chosen as previous in cases of adjunct clauses does result in fewer violations of Strong C1 than with Kameyama's, but not so many that C1 is verified, and only if relative clauses are treated as utterances. And in case the adjunct clause comes at the beginning of a sentence, as in *if*-clauses, it is best to follow the linear sequence rather than treating it as embedded.

[95] 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 et al. (1993) and Passonneau (1993). 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.

Minimizing violations should not be the overriding goal We already said in Section §3 that we see the primary function of our results to clarify the extent to which real texts are governed by Centering's preferences, rather than to suggest how Centering's parameters should be set. We don't think that minimizing violations should be the only factor taken into account when deciding how to set parameters. (Remember that in previous sections we argued against treating PRO2s as realizations of CFs, on the (linguistic) grounds that the referents of PRO2s are best viewed as being interpreted deictically, even though doing this would result in fewer violations of Strong C1; we also argued against having all NPs in predicative position introduce CFs, as this would lead to rather implausible semantic claims.) Some violations are best accepted, and explained in terms of the interaction of Centering preferences with other preferences. (Some examples of how such an interaction might be formalized are discussed below.)

Special care is needed when alternative definitions are supported by cross-linguistic evidence, or by the results of psychological studies. Among the parameters of the theory whose setting cannot be determined only by our results are the choice of ranking function, whether the ranking function should be partial or total, and the definition of utterance (finite clauses, sentences, or a more 'semantic' definition?). 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.[96] 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 et al. 1986; Gordon et al. 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 et al. 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 (see, e.g., (Traum and Heeman 1997)).

7.2 The claims of Centering, revisited

Having reached the conclusion that it is possible to find parameter instantiations under which all three claims are verified–provided that we interpret them as preferences, verified in a statistical sense–we will now examine the three claims individually.

[96] It would be interesting to compare STRUBE-HAHN with GFTHERELIN in Hindi.

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 et al. 1995, 1983)–are very robust. Rule 1 (GJW 95) and Rule 1 (GJW 83) are verified with all parameter instantiations, and in a very convincing way: in the instantiations we considered, the percentage of violations of Rule 1 (GJW 95) never exceeds 8% of the total number of utterances.

On the other hand, one should keep in mind that these two versions of Rule 1 make very weak claims about pronominalization. All that Rule 1 (GJW 95) says is that *if* we decide to pronominalize, *then* we should pronominalize the CB. This formulation doesn't address the real problem for a theory of pronominalization or, more in general, of NP form decision, which is to decide when a discourse entity should be realized as a pronoun (Henschel et al. 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 et al. 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 (22) 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.[97] 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.[98]

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%); and this even when a pretty fine-grained notion of segment is used. 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 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[99] expressed by texts in the museum domain of our

[97] 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).

[98] This view of the process of choosing the realization of a discourse entity as a balancing act was also put forward by Gundel et al. (1993) and Almor (1999). The discrepancy between pronominalization and CB-hood in our corpus is analyzed in more detail by Henschel et al. (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.

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

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.[100]

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 (45), repeated here:

(45) (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.

In (46) we do not find explicit discourse connectives, but the discourse is quite clearly coherent in virtue of the fact that each of the items in the list specifies a possible counterindication to the use of the medicine in question. In such cases, treating second person pronouns as CF realizations would also result in the utterances being classified as 'entity-coherent,' but explaining local coherence in terms of rhetorical relation does not require making such dubious assumptions.

(46) Are you sensitive or allergic to any oestrogens? Are you sensitive or allergic to any of the inactive ingredients? Are you pregnant, planning a pregnancy or think you may be pregnant. Are you breast feeding? Do you have, or have you ever had, cancer of the breast or uterus? Have you experienced any unusual vaginal bleeding recently?

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 generic point that will then be illustrated by a specific object. In (47), (u2) gives background concerning the decoration of a cabinet. In (48), utterances (u2)-(u5) give information about a particular class of rings to which the objects under discussion belong.

(47) (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.

(48) (u1) Two gold finger-rings from Roman Britain (2nd - 3rd century AD).

(u2) Polygonal openwork rings incorporating an inscription are a distinctive type found throughout the Empire. (u3) The pierced technique is especially typical of late Antique jewelry, (u4) but this class of ring appears to have come into use in the 2nd century AD. (u5) In many cases the mottoes on the

[100] Furthermore, as Karamanis and Manurung (2002) had already argued, given the combinatorially exponential number of alternative text plans, a heuristic strategy not always resulting in the 'best' text plan according to a given evaluation function (such as minimizing the number of violations of entity coherence) may be the only feasible way of arriving at a text plan.

panels are in Greek: That on 602 (left), from Corbridge, Northumberland, reads: 'the love-token of Polemios'.

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. Whereas in the case of (u1)-(u2) in (48) *poligonal openwork rings* may be conceivably viewed as a bridging reference to *two gold finger-rings*, it is more difficult to find a clear bridging reference in (u3).[101] While it is true that some of these violations could be fixed by adopting a broader notion of bridging reference–e.g., in (47) 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).[102]

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.

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 (if not the two 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.[103] The version of Rule 2 from (Grosz et al. 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

[101] Of course, identifying utterances with sentences would also ensure the presence of a link by merging together several clauses.

[102] Other sources of coherence are possible as well. E.g., Karamanis (2001) examines the possibility that TEMPORAL FOCUSING (Webber 1988; Kameyama et al. 1993) may be the main source of coherence in certain texts, such as narratives. 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).

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

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 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. However, we found in Section §5 that some linguistic predictions based on transitions only hold, or are much stronger, if EST and CON are not conflated. For example, we found that the hypothesis that pronouns in subject position somehow suggest a continuation only holds if we consider EST as a type of shift rather than as a continuation.

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. Kibble's 'decomposition' of Rule 2 is a good way of looking at which of its underlying 'cohesive principles' is verified most frequently. As we saw, the 'Kibble score' changes rather dramatically from version to version, to reach its highest value (2.21) with indirect realization, u=s, and Strube and Hahn ranking. In this version, more than 2 / 3 of utterances are continuous; on the other hand, less than 2 / 7 are cheap, salient or cohesive.[104]

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 et al. 1993)

7.3 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. In this section we discuss a few of these.

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 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

[104] Karamanis (2001) argues that the four principles proposed by Kibble should not be seen as additive, but as 'weighted' in the sense of Optimality Theory.

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; this is very common with relative clauses, as in the following example, from the *Guardian* newspaper:

(49) 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 et al. 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 2003), 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. $\text{CF}(\text{U}_{i-1}) \cap \text{CF}(\text{U}_i) \neq \emptyset$;
2. or there is a rhetorical relation **RR** such that **RR**(U_{i-1},U_i);[105]

[105]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.

3. or U_{i-1} and U_i are temporally coherent in the sense, e.g., of (Kameyama et al. 1993);
4. ... (other)

Oberlander and Poesio (2002) explore a stronger formulation, requiring the existence of an *explicit* link between the two utterances: either an explicit anaphoric reference with a limited range of relations, or an explicit rhetorical connective. 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 (20) 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 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 et al. 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. Continuations of these sentences in which both entites are pronominalized, like (50b), seem to us to be more felicitous than continuations in which only one or neither is, like (50c-e), especially when the continuation is not a separate sentence.

(50) a. Mary took a nickel from her toy bank yesterday, and
b. she put it on the table near Bob.
c. she put the nickel on the table near Bob.
d. Mary put it on the table near Bob.
e. Mary put the nickel on the table near Bob.

Further evidence against the move of forcing the ranking function to be partial to avoid utterances having more than one CB comes from Japanese data discussed by Hara (2003). In Centering-based theorizing on topicality and pronominalization in Japanese, it is generally assumed that *wa*-marked NPs identify the CB; indeed, Walker et al. (1994) argue that "the use of *wa* in a discourse-initial utterance instantiates the *wa*-marked entity as the CB" ((Walker et al. 1994), section 4.1). However, example (51)–from Miyuki Miyabe's "Brave Story", discussed in (Hara 2003)–shows that in contexts like (20), *both* the animate entity (Akira Miyoshi) and the inanimate one (the steel company) can be *wa*-marked, as in (51b), although in these cases one of the *wa*-marks is typically taken to have a contrastive use.

(51) a. Wataru-no chichi-no Miyoshi Akira-wa seitetsu-gaisha-ni
Wataru-gen father-gen Miyoshi Akira-top steel company-dat
tsutomete iru
working-be
"Akira Miyoshi, Wataru's father, is working for a steel company"

b. Akira-wa seitetsu-no genba-ni-wa tankikan-sika
Akira-top iron-manufacturing-gen factory-dat-top short-period-only
ita koto-ga nai
was fact-nom not
"Akira was at the iron works for a short period only"

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 et al. 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, (20), neither Branicki nor the cupboard occur in u229 in a position that would be subject to RNP effects according to Gordon *et al.*. In other words, (some) evidence used by Grosz *et al.* in support of CB uniqueness cannot be used to argue that u229 in (20) has a single CB. This evidence is also consistent with a different solution of the problem raised by examples like (20): 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.[106]

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*

[106]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.

al.– roughly, the finding that using a proper name in subject position to refer to an entity also realized by a proper name in subject position in the previous sentence results in slower reading times– might be an instance of this more general phenomenon.

7.4 Limitations of this study

We conclude by listing a few shortcomings of this work that we would like to be addressed in future investigations.

Other domains The major limitation of this study is that it concentrated on only two genres. It would be useful to perform a similar analysis with texts from genres such as narratives and dialogues. This said, we would like to emphasize that at least one of the domains under study, that of museum descriptions, ought to be ideally suited for a theory of entity coherence, in that most texts are about objects and their relationships to other objects.

Semantic structuring A second limitation of this work is that it concentrated on the effect of syntactic factors on salience; it would be useful to study the impact of semantic factors such as thematic roles, when we know how to annotate them reliably. The study of the impact of rhetorical structure in Section §6 is a first step in this direction.

Bridging It is obviously the case that with a more thorough annotation of bridging references one would get fewer violations of C1. The difficult question is whether it is possible to do so in a reliable fashion.

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, 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. Hua Cheng was in part supported by the EPSRC project GNOME, GR/L51126/01. 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

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, 106:748–765.

Alshawi, H. (1987). *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge.

Anderson, A., Garrod, S., and Sanford, A. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *Quarterly Journal of Experimental Psychology*, 35:427–440.

André, E., Poesio, M., and Rieser, H., editors (1999). *Proc. of the ESSLLI Workshop on Deixis, Demonstration and Deictic Belief in Multimedia Contexts*, Utrecht. University of Utrecht.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. Croom Helm Linguistics Series. Routledge.

Arnold, J. E. (1998). *Reference Form and Discourse Patterns*. PhD thesis, Stanford University.

Asher, N. (1993). *Reference to Abstract Objects in English*. D. Reidel, Dordrecht.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proc. of the 36th ACL*.

Barker, C. (1991). *Possessive Descriptions*. PhD thesis, University of California at Santa Cruz, Santa Cruz, CA.

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

Brennan, S., Friedman, M., and Pollard, C. (1987). A centering approach to pronouns. In *Proc. of the 25th ACL*, pages 155–162.

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

Brennan, S. E. (1998). Centering as a psychological resource for achieving joint reference in spontaneous discourse. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering in Discourse*, chapter 12, pages 227–249. Oxford University Press.

Byron, D. and Stent, A. (1998). A preliminary model of centering in dialog. In *Proc. of the 36th ACL*.

Caramazza, A., Grober, E., Garvey, C., and Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior*, 16:601–609.

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

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

Cheng, H., Poesio, M., Henschel, R., and Mellish, C. (2001). Corpus-based NP modifier generation. In *Proc. of the Second NAACL*, Pittsburgh.

Chinchor, N. A. and Sundheim, B. (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, H. H. (1977). Bridging. In Johnson-Laird, P. N. and Wason, P., editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, London and New York.

Cooreman, A. and Sanford, T. (1996). Focus and syntactic subordination in discourse. Research Paper RP-79, University of Edinburgh, HCRC.

Corbett, A. and Chang, F. (1983). Pronoun disambiguating: Accessing potential antecedents. *Memory and Cognition*, 11:283–294.

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

Dahl, O. and Fraurud, K. (1996). Animacy in grammar and discourse. In Fretheim, T. and Gundel, J. K., editors, *Reference and Referent Accessibility*, number 38 in Pragmatics and Beyond, pages 47–64. John Benjamins.

Dale, R. (1992). *Generating Referring Expressions*. The MIT Press, Cambridge, MA.

Di Eugenio, B. (1998). Centering in italian. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering Theory in Discourse*, chapter 7, pages 115–138. Oxford.

Di Eugenio, B., Moore, J. D., and Paolucci, M. (1997). Learning features that predict cue usage. In *Proc. of the 35th ACL*, Madrid.

Fox, B. A. (1987). *Discourse Structure and Anaphora*. Cambridge University Press, Cambridge, UK.

Garrod, S. and Sanford, A. J. (1994). Resolving sentences in a discourse context. In Gernsbacher, M. A., editor, *Handbook of Psycholinguistics*, chapter 20, pages 675–698. Academic Press.

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

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

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

Gordon, P. C. and Chan, D. (1995). Pronouns, passives and discourse coherence. *Journal of Memory and Language*, 34:216–231.

Gordon, P. C., Grosz, B. J., and Gillion, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–348.

Gordon, P. C. and Hendrick, R. (1999). The representation and processing of coreference in discourse. *Cognitive Science*, 22:389–424.

Gordon, P. C., Hendrick, R., Ledoux, K., and Yang, C. L. (1999). Processing of reference and the structure of language: an analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.

Gordon, P. C. and Scearce, K. A. (1995). Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory and Cognition*, 23:313–323.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics, vol.3: Speech Acts*, pages 41–58. Academic Press, New York.

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

Grosz, B., Joshi, A., and Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proc. ACL-83*, pages 44–50.

Grosz, B. J. (1977). *The Representation and Use of Focus in Dialogue Understanding*. PhD thesis, Stanford University.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1986). Towards a computational theory of discourse interpretation. Unpublished ms.

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

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

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

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

Hahn, U. and Strube, M. (1997). Centering in-the-large: Computing referential discourse segments. In *Proc. of the 35th Meeting of the ACL*, Madrid.

Hara, Y. (2003). A Japanese example of Centering transitions: Zero topic and grammatical topic. Unpublished manuscript.

Hawkins, J. A. (1978). *Definiteness and Indefiniteness*. Croom Helm, London.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts at Amherst.

Henschel, R., Cheng, H., and Poesio, M. (2000). Pronominalization revisited. In *Proc. of 18th COLING*, Saarbruecken.

Hitzeman, J., Black, A., Taylor, P., Mellish, C., and Oberlander, J. (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),*, page Paper 591, Australia.

Hitzeman, J., Mellish, C., and Oberlander, J. (1997). Dynamic generation of museum web pages: The intelligent labelling explorer'. *Journal of Archives and Museum Informatics*, 11:107–115.

Hitzeman, J. and Poesio, M. (1998). Long-distance pronominalisation and global focus. In *Proc. of ACL/COLING, vol. 1*, pages 550–556, Montreal.

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

Hudson, S. B., Tanenhaus, M. K., and Dell, G. S. (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, S. and Tanenhaus, M. K. (1998). Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering in Discourse*, pages 199–226. Oxford University Press.

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

Iida, M. (1998). Discourse coherence and shifting centers in japanese texts. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering Theory in Discourse*, chapter 9, pages 161–180. Oxford University Press.

Jarvella, R. J. and Klein, W., editors (1982). *Speech, Place and Action - Studies in Deixis and Related Topics*. John Wiley, Chichester and New York.

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

Joshi, A. K. and Weinstein, S. (1981). Control of inference: Role of some aspects of discourse structure–centering. In *Proc. International Joint Conference on Artificial Intelligence*, pages 435–439.

Kameyama, M. (1985). *Zero Anaphora: The case of Japanese*. PhD thesis, Stanford University, Stanford, CA.

Kameyama, M. (1986). A property-sharing constraint in centering. In *Proc. ACL-86*, pages 200–206.

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

Kameyama, M., Passonneau, R., and Poesio, M. (1993). Temporal centering. In *Proc. of the 31st ACL*, pages 70–77, Columbus, OH.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

Kaplan, D. (1979). Dthat. In Cole, P., editor, *Syntax and Semantics v. 9, Pragmatics*, pages 221–243. Academic Press, New York.

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

Karamanis, N. (2003). *Entity coherence for descriptive text structuring*. PhD thesis, University of Edinburgh, Informatics.

Karamanis, N. and Manurung, H. M. (2002). Stochastic text structuring using the principle of continuity. In *Proc. of INLG*, pages 81–88.

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

Kehler, A. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3).

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

Kibble, R. (2001). A reformulation of Rule 2 of Centering Theory. *Computational Linguistics*, 27(4):579–587.

Kibble, R. and Power, R. (2000). An integrated framework for text planning and pronominalization. In *Proc. of the International Conference on Natural Language Generation (INLG)*, Israel.

Kintsch, W. and van Dijk, T. (1978). Towards a model of discourse comprehension and production. *Psychological Review*, 85:363–394.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text representation: linguistic and psycholinguistic aspects*. John Benjamins.

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

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

Lesgold, A., Lajoie, S., Bunzo, M., and Eggan, G. (1992). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In Larkin, J. and Chabay, R., editors, *Computer assisted instruction and intelligent tutoring systems: Shared issues and complementary approaches*, pages 201–238. Erlbaum, Hillsdale, NJ.

Linde, C. (1979). Focus of attention and the choice of pronouns in discourse. In Givon, T., editor, *Syntax and Semantics 12*. Academic Press.

Loebner, S. (1987). Definites. *Journal of Semantics*, 4:279–326.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, 8(3):243–281.

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

Marcu, D., Romera, M., and Amorrortu, E. (1999). Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. In *Workshop on Levels of Representation in Discourse*, pages 71–78. University of Edinburgh.

May, R. (1977). *The Grammar of Quantification*. PhD thesis, MIT, Cambridge, MA.

May, R. (1985). *Logical Form in Natural Language*. The MIT Press.

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

Miltsakaki, E. (1999). Locating topics in text processing. In *Proc. of CLIN*.

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

Moore, J. and Pollack, M. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Moser, M. and Moore, J. D. (1996a). On the correlation of cues with discourse structure: Results from a corpus study. Unpublished manuscript.

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

Moser, M., Moore, J. D., and Glendening, E. (1996). Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units. Technical Report 96-17, University of Pittsburgh, Department of Computer Science.

Oberlander, J., O'Donnell, M., Knott, A., and Mellish, C. (1998). Conversation in the museum: Experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Hypermedia and Multimedia*, 4:11–32.

Oberlander, J. and Poesio, M. (2002). Entity coherence and relational coherence: a corpus-based investigation. Presented at the Berlin workshop on Topics in Discourse.

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

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

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

Passonneau, R. J. and Litman, D. (1993). Feasibility of automated discourse segmentation. In *Proceedings of 31st Annual Meeting of the ACL*.

Pearson, J., Stevenson, R., and Poesio, M. (2000). Pronoun resolution in complex sentences. In *Proc. of AMLAP*, Leiden.

Pearson, J., Stevenson, R., and Poesio, M. (2001). The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In Poesio, M., editor, *Proc. of the First Workshop on Cognitively Plausible Models of Semantic Processing (SEMPRO)*. University of Edinburgh, HCRC.

Poesio, M. (1994). Weak definites. In Harvey, M. and Santelmann, L., editors, *Proceedings of the Fourth Conference on Semantics and Linguistic Theory, SALT-4*, pages 282–299. Cornell University Press.

Poesio, M. (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.

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

Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*, Boston.

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

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J. M., Kibble, R., and Stevenson, R. (2000). Specifying the parameters of Centering Theory: a corpus-based evaluation using text from application-oriented domains. In *Proc. of the 38th ACL*, Hong Kong.

Poesio, M. and Di Eugenio, B. (2001). Discourse structure and anaphoric accessibility. In Kruijff-Korbayová, I. and Steedman, M., editors, *Proc. of the ESSLLI 2001 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

Poesio, M. and Nygren-Modjeska, N. (2002). The THIS-NP hypothesis: A corpus-based investigation. In *Proc. of DAARC*, Lisbon.

Poesio, M., Patel, A., and Di Eugenio, B. (2004). Discourse structure and anaphora. In preparation.

Poesio, M. and Stevenson, R. (To appear). *Salience: Theoretical Models and Empirical Evidence*. Cambridge University Press, Cambridge and New York.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.

Portner, P. H. and Yabushita, K. (1998). The semantics and pragmatics of topic phrases. *Linguistics and Philosophy*, 21:117–157.

Prasad, R. and Strube, M. (2000). Discourse salience and pronoun resolution in Hindi. In *Penn Working Papers in Linguistics*, volume 6, pages 189–208.

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

Prince, E. F. (1992). The ZPG letter: subjects, definiteness, and information status. In Thompson, S. and Mann, W., editors, *Discourse description: diverse analyses of a fund-raising text*, pages 295–325. John Benjamins.

Prince, E. F. (1998). Subject-prodrop in yiddish. In Bosch, P. and van der Sandt, R., editors, *Focus: linguistic, cognitive, and computational perspective*, pages 82–104. Cambridge.

Quirk, R. and Greenbaum, S. (1973). *A University Grammar of English*. Longman, Harlow, Essex, England.

Rambow, O. (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).

Reichman, R. (1985). *Getting Computers to Talk Like You and Me*. The MIT Press, Cambridge, MA.

Reinhart, T. (1983). *Anaphora and semantic interpretation.* Croom Helm, London.

Sanford, A. J. and Garrod, S. C. (1981). *Understanding Written Language*. Wiley, Chichester.

Scott, D., Power, R., and Evans, R. (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, P. (1967). Functional sentence perspective in a generative description. *Prague Studies in Mathematical Linguistics*, 2:203–225.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in English discourse*. PhD thesis, MIT.

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

Stevenson, R., Knott, A., Oberlander, J., and McDonald, S. (2000). Interpreting pronouns and connectives: interactions between focusing, thematic roles and coherence relations. *Language and Cognitive Processes*, 15.

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus, and the representation of events. *Language and Cognitive Processes*, 9:519–548.

Strube, M. (1998). Never look back: An alternative to centering. In *Proc. of COLING-ACL*, pages 1251–1257, Montreal.

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

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

Szabolcsi, A., editor (1997). *Ways of Scope Taking*. Kluwer, Dordrecht.

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

Traum, D. and Heeman, P. (1997). Utterance units in spoken dialogue. In Maier, E., Mast, M., and Luperfoy, S., editors, *Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence, pages 125–140. Spring-Verlag.

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

Vallduvi, E. (1990). *The Informational Component*. PhD thesis, University of Pennsylvania, Philadelphia.

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

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

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

Walker, M. A. (1996). Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264.

Walker, M. A. (1998). Centering, anaphora resolution, and discourse structure. In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, *Centering in Discourse*, chapter 19, pages 401–435. Oxford University Press.

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

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

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

Webber, B. L. (1978). A formal approach to discourse anaphora. Report 3761, BBN, Cambridge, MA.

Webber, B. L. (1988). Tense as discourse anaphor. *Computational Linguistics*, 14(2):61–73.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Woods, A., Fletcher, P., and Hughes, A. (1986). *Statistics in Language Studies*. Cambridge University Press.