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

A Corpus-Based Evaluation of Centering Theory

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

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

Hua Cheng
BCL Computers Inc
hua@bcl-computers.com

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

CS Technical Report CSM-369

University of Essex, Department of Computer Science, NLE Group

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

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

A Corpus-Based Evaluation of Centering Theory

Abstract

Centering Theory has proven a 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, e.g., (zero) pronominalization in several languages. But because these concepts were intentionally left underspecified in the formulation of the theory, the claims made in the framework have only ever been tested by fixing upon a particular way of instantiating the theory's 'parameters' - e.g., by assuming that ranking is based on grammatical function. This leaves open the question of whether all the particular choices in a certain study were optimal. Furthermore, none of the previous corpus-based studies checked that the annotation only involved notions that could be annotated reliably.

We tested the claims of the theory in a more general way than in these previous studies, by trying to find the best instantiation for English of the 'parameters' of the theory among those proposed in the literature, as well as the version of the claims with the fewest violations. We did this by annotating a corpus of texts from two distinct domains with information that has been claimed to affect the instantiation of the parameters of the theory and can be annotated in a reliable fashion. We found, first, that at least two versions of Rule 1 are verified under most ways of computing the basic notions of the theory. Second, that the validity of Constraint 1 depends more than that of Rule 1 on the way parameters of the theory such as utterance and realization are defined; in particular, if we identify utterances with finite clauses, and only allow for direct realization, significantly more utterances violate the best-known version of the constraint than satisfy it. Third, that Rule 2 is very sensitive to the definitions, being verified only under a very few ways of specifying the parameters. We also found that it is very difficult to say which is the 'best' way of specifying these parameters, since there is a tradeoff between Constraint 1 and Rule 1 - trying to reduce the number of utterances without a backward-looking center (CB) results in an increased number of cases in which some discourse entity, but not the CB, gets pronominalized, and viceversa.

1 MOTIVATIONS

Centering Theory (Joshi and Weinstein 1981; Grosz et al. 1983, 1995; Walker et al. 1998b) is the part of Grosz and Sidner's general 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 discourse segments. A fundamental characteristic of Centering Theory, and a key difference from earlier theories of focusing more directly concerned with anaphora resolution such as Sidner's (1979), is that it is more of a *linguistic* theory - and a very abstract one - than a computational theory. By this we mean, first of all, that its primary aim is to make cross-linguistically valid claims about (certain aspects of) discourse viewed as a linguistic phenomenon, rather than to provide algo-

rithms for anaphora resolution or anaphora generation (although a variety of algorithms based on the theory have been proposed in the literature). 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 theory very different from those typically proposed in the literature on anaphora in the fields of Natural Language Processing and Natural Language Generation; indeed, one that leaves many computational linguists disconcerted. Particularly disturbing is the fact that papers such as (Grosz et al. 1995) do not specify algorithms for computing central notions of the theory such as 'utterance', 'previous utterance', 'ranking,' and 'realization'. ((Grosz et al. 1995) claim that while these concepts play a central role in any theory of discourse coherence and salience, their precise characterization should be left for subsequent research; indeed, notions such as ranking might be defined in a different way for each language (Walker et al. 1994).) In fact, different definitions of the central notions of the theory have been proposed, often by the same authors (cfr., e.g., the different definitions of CB in (Grosz et al. 1983, 1995; Gordon et al. 1993). This situation is very similar to the one we encounter in syntax, where new definitions of 'command' are continuously being proposed. But it makes Centering rather different from the theories most commonly encountered in computational linguistics, which tend to involve detailed specifications of algorithms (Hobbs 1978; Sidner 1979; Carter 1987; Lappin and Leass 1994; Hardt 1997; Mitkov 1998; Vieira and Poesio 2000).

This underspecification doesn't mean however that Centering is merely an attempt at mapping out a field of research, without making any specific claims. On the contrary, sometimes the theory's claims are stronger than those of competing theories, as we will see below. It does mean though that two theories 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 versions, many people wonder 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 work presented in this paper had two main goals: to find out the extent to which 'core' claims of Centering such as Constraint 1, Rule 1, and Rule 2 (introduced below) are actually verified, i.e., how many violations of these claims one actually encounters, under several ways of instantiating the parameters of the framework, compared on the same data. In doing so, we also intended to find out which of the many way of specifying the 'parameters' of Centering would make these claims most accurate as predictors of coherence and pronominalization for English. We did this by annotating a corpus of English texts with the sort of information required to implement a number of variants of Centering Theory, and using this corpus to automatically check the claims under a variety of 'parameter configurations'. Other corpus-based studies of Centering already exist (Walker 1989; Passonneau 1993; Byron and Stent 1998; Di Eugenio 1998; Kameyama 1998; Strube and Hahn 1999; Tetreault 1999), but they only compared two or three instantiations of the Centering framework; the present study is more systematic than these earlier studies both in that it considers a greater number of parameters, and more possible parameter configurations; and in that we carefully checked our annotation techniques to ensure reliability. We also developed an annotation manual that can be used to extend

our analysis to other data.

Another difference from earlier comparisons between versions of Centering Theory is that we wanted to evaluate the predictions of the theory in domains of interest for real applications–Natural Language Generation, in our case. For this reason, we used texts in two genres not yet studied in the Centering literature, but of interest to developers of NLG systems: instructional texts and descriptions of museum objects to be displayed on Web pages.

The paper is organized as follows. We first review the basic notions of the theory. We then discuss how the corpus was annotated, and how the annotation was used. In Section 4 we present our main results. In Section 5 we re-examine a few claims made in the Centering literature concerning the linguistic impact of notions from Centering. In the following section we briefly report a second series of experiments investigating the impact of rhetorical structure. The final discussion is in Section 6.

2 FUNDAMENTALS OF CENTERING THEORY

Intuitions

Centering Theory is simultaneously a theory of discourse coherence and of discourse salience. As a theory of coherence, it attempts to characterize ENTITY-COHERENT discourses - i.e., discourses that can be considered coherent on the basis of the way discourse entities ('topics') are introduced and discussed.[1] At the same time, it is also intended to be a theory of *salience*: i.e., to predict which entities will be most salient at any given time - which makes it also a theory of pronominalization as well, given that the entities most salient are those most likely to be pronominalized (Grosz et al. 1995; Gundel et al. 1993). We can't include in this paper a full description of, and motivation for, the many versions of Centering proposed in the literature; we simply summarize in this section these existing proposals in enough detail to allow the reader to follow the subsequent discussion, and refer the reader to basic references such as (Grosz et al. 1995), the papers in (Walker et al. 1998b), or the discussion in (Poesio and Stevenson pear).[2]

The main claim of the theory is that every utterance in a discourse has a 'central entity', called the 'Backward-Looking Center', or CB. A second, and almost equally important, claim is that the discourse entities 'realized' by an utterance (more on the notion of 'realization' below) are *ranked*, and the identity of the CB is crucially determined by this ranking. The primary intuition about local coherence that Grosz *et al.* attempt to capture is the idea that discourses in which successive 'utterances' keep mentioning the same discourse entities are perceived as 'more coherent' than discourses in which different entities are mentioned, already advanced in work such as (Chafe 1976; Kintsch and van Dijk 1978; Givon 1983). One role of the CB is, then, to 'link back' the utterance in which it occurs to the previous discourse. As far as local salience is concerned, Centering Theory's basic contention is that the CB is most likely to be realized as a pronoun than other entities. These views, as well, are shared with a number of other theories of discourse, whether in the psychological (e.g., (Sanford and Garrod 1981)), computational (Alshawi 1987; Sidner 1979) or linguistic literature (Prince 1981;

[1] Entity-based theories of coherence are so-called by contrast with relation-centered theories of coherence, such as those developed in (Hobbs 1979; Mann and Thompson 1987) and used in (Reichman 1985; Fox 1987; Lascarides and Asher 1993). The earliest detailed entity-based theory of coherence we are aware of is the one proposed in Kintsch and van Dijk (1978), who also explicitly mention the need to supplement such theories with a theory of relational coherence. We return on the topic below; further discussion of 'entity-centered' vs. 'relation-centered' notions of coherence is in (Knott et al. ress; Stevenson et al. 2000).

[2] Among the important papers in the field not included in Walker *et al.*'s collection we should mention at least (Kameyama 1985; Brennan et al. 1987; Passonneau 1993; Brennan 1995; Kehler 1997; Strube and Hahn 1999; Kibble pear).

Givon 1983); the claim that pronominalization correlates with higher ranking is especially explicit in (Ariel 1990; Gundel et al. 1993).

According to Centering Theory, the ranking of CFs is determined by factors such as the grammatical function of the NP realizing the discourse entity, the discourse status of the discourse entity and in particular whether it has been pronominalized before (Grosz et al. 1995, p. 212). (The factors affecting ranking are discussed below.) This claim is motivated by the contrast between examples like (1) and (2).[3]

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

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

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

In fact, Grosz *et al.* go much further than this, arguing (against (Sidner 1979), among others) that only one discourse entity is the 'center' in each utterance. As evidence in support of this claim, they note the contrast between continuations (c)-(f) of the discourse initiated by utterances (3a-b) (these are examples (7) through (10), Grosz, Joshi and Weinstein, 1995, p. 211-212).

(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.

According to Grosz et al, there is a marked decrease in acceptability from (3c) to (3f)), whereas if both Susan and Betsy were equally highly ranked after utterance (b), all variants would be equally acceptable.

According to Grosz et al, ranking affects coherence as well, in the sense that the fewer the changes in the ranking of discourse entities across utterances, the more a text feels (locally) coherent. This claim is motivated by the contrast between examples like (4) and (5) (these are examples (7) and (8) from (Grosz et al, 1995)):

(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.

[3](1) is GJW's (15), p. 215; these examples are also discussed by Kehler (1997).

(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* 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.

Concepts and Definitions

The fundamental assumption of Centering is that when processing a discourse, a local attentional state, or LOCAL FOCUS, is continuously updated; the minimal update unit is called the UTTERANCE. The local focus consists of a set of FORWARD-LOOKING CENTERS (CFs), which might be thought of as mentions of discourse entities (Karttunen 1976; Webber 1978; Heim 1982; Kamp and Reyle 1993) or 'potential discourse foci' (Sidner 1979) in a given utterance. The local focus also contains information about the relative prominence of these CFs, some of which are especially singled out. Utterances update the local focus by replacing the current (CFs) with new ones. 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) formalize the relationship between utterances and CFs by means of one of their so-called 'Constraints':[4]

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

We saw above that an important claim of the theory is that forward-looking centers are RANKED, and that because of this ranking, some CFs acquire particular prominence. The most highly ranked CF realized by an utterance (when one exists) is called the 'Preferred Center', or CP. Furthermore, and most importantly, the notion of ranking is used to characterize one of the CFs as the BACKWARD-LOOKING CENTER (CB). The CB is the closest concept in Centering Theory to the traditional notion of 'topic' (Sgall 1967; Chafe 1976; Givon 1983; Reichman 1985; Vallduvi 1990) and plays a central role in the theory's claims about both local coherence and local salience (discussed below).

Although in the original paper (Grosz et al. 1983) the CB was only characterized in intuitive terms, subsequent work within the framework has been usually based on the following definition of the CB of utterance U_i in terms of ranking, proposed by Grosz et al. (1995) and 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.

[4]The theory as formulated by Brennan et al includes two more 'constraints', discussed below. It should be noted that these three Constraints 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.

Notice that according to this definition the value of the CB depends exclusively on ranking and 'previous utterance', thus making the specification of these notions crucial to the predictions of a particular implementation of the framework. This is in clear contrast with 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. We will consider competing characterizations of the CB below.

Finally, the intuition that texts are perceived to be (locally) more coherent when successive utterances are packaged in such a way as to be perceived as being 'about' the same discourse entity has been formalized in the Centering model 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: i.e., whether the CB and the CP change. Many such classifications of transitions have been proposed. (Grosz et al. 1995) distinguished 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) is also the most highly ranked entity (CP) of that utterance:

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

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

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

We will review a few alternative classifications below.

Main Claims

Centering Theory is not simply a conceptual vocabulary; the notions introduced above are used to formulate the three main claims of the theory, for which we follow again the terminology of 'constraints' and 'Rules' proposed by Brennan et al:

Constraint 1 (Strong): All utterances of a segment except for the 1st have exactly one CB.[5]

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

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

Constraint 1 is, first of all, a claim about local coherence: namely, that there is a preference in discourses to continue talking about the same entities. If one sees the notion of CB as a formalization of the idea of 'topic' (Gundel 1998; Hurewitz 1998; Miltsakaki 1999) , the constraint can be seen as a claim that there is exactly one (or no more than one) 'topic' at each point. (Whereas, say, Sidner's theory can be seen as involving two 'topics', and theories such as (Givon 1983; Alshawi 1987; Lappin and Leass 1994; Arnold 1998) assume that being a 'topic' is only a matter of degree - alternatively, that there can be an arbitrary number of topics.) Constraint 1 also makes a claim about salience, in the sense of 'likelihood to be pronominalized' - see, e.g., (Gundel et al. 1993): that there is exactly one most salient entity at each point. Rule 1 is the main claim of the theory about pronominalization, stating a preference for pronominalizing the CB, if anything is pronominalized at all. Rule 2 is,

[5] As we will see below, a weaker variant of this Constraint has also been proposed (Walker et al. 1998a), according to which all utterances have *at most one* CB.

again, a claim about coherence: 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.

More generally, these claims express what is perhaps the most distinguishing feature of Grosz and Sidner's general 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 (where, according to Grosz and Sidner, the attentional state is 'parasitic on the intentional level' which ensures (global) coherence), and at the local level, where the data structure whose values determine whether a text is perceived as being coherent also controls pronominalization. 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).

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)

The Parameters of Centering Theory

The concepts introduced in Centering to theorize about the local focus - 'utterance', 'previous utterance', 'ranking,' and 'realization' - were left essentially undefined by Grosz, Joshi and Weinstein, although they suggested ways of defining them. Similarly undefined is the notion of 'pronoun' governed by Rule 1: should it include only third person singular pronouns? Or also plural ones? What about second person pronouns? Without providing full specifications of such notions it is impossible to evaluate the claims above - just as, say, the predictions of Government and Binding theory cannot be tested without giving a fully explicit definition of 'command' or 'argument'. As a result, a considerable amount of work in the area has been concerned with establishing the best 'parameter instantiations': we review some of these proposals here.[6]

Utterance and Previous Utterance In the early Centering papers, utterances were implicitly identified with 'sentences'. Kameyama (1998), however, noted that such identification makes anaphoric expressions much more ambiguous than if they were resolved clause by clause and, furthermore, it leads to problems with multiclausal sentences: e.g., grammatical function ranking becomes difficult to compute, as a sentence may have more than one subject. Kameyama proposed instead that the local focus is updated by each tensed clause, rather than by each sentence; and she classified finite clauses into (i) utterance units that constitute a 'permanent' update of the local focus, such as coordinated clauses and adjuncts, and (ii) utterance units that result in temporary updates that are then 'popped', much as the information introduced into discourse by subordinated discourse segments. Kameyama called units of this second type EMBEDDED utterance units, and proposed that clauses that serve as adjuncts, or as complements of certain verbs, behave this way. For example, Kameyama proposes to break up (6) as follows:

(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

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

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) led them to propose that some types of adjuncts–in particular, clauses headed by *after* and *before*–should be treated as 'embedded' rather than as '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) 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; this hypothesis wasn't however tested. Other subordinate clauses and parentheticals are not discussed by either Kameyama or Suri and McCoy.

Kameyama's identification of utterances with tensed or finite clauses has recently been questioned in work such as (Strube and Hahn 1998; Miltsakaki 1999). Miltsakaki (1999) argues, on the basis of cross-linguistic data from English and Greek, that utterances are best identified with sentences, and that only the main clause should be considered for the ranking.

Realization Grosz et al. (1995) consider two possible 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 CF. INDIRECT realization is when one of the noun phrases in the utterance is a bridging reference to that CF in the sense of (Clark 1977), 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) John walked towards the house. 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 the second utterance by virtue of being referred to by the bridging reference *the door* (see, e.g., the discussion in (Grosz et al. 1995; Walker et al. 1998b)).

Ranking Perhaps the most discussed parameter in Centering –at least in the versions of the theory that accept the definition of CB specified by Constraint 3– is ranking. All theories based on Centering 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 ranking factors may not be the same in all languages. Nevertheless, at least as far as English is concerned, most versions of the theory ever 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. Specifically, (Grosz et al. 1995) claim that subjects are ranked more highly than objects, and these are ranked more highly than other grammatical positions - summarized as 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 and indirect objects), and by Walker et al. (1994); Turan (1995) for Japanese and Turkish, respectively. There is quite a lot of psychological support for at least the idea that subjects are more highly ranked (Hudson et al. 1986; Gordon et al. 1993; Brennan 1995; Hudson-D'Zmura and Tanenhaus 1998).

In more recent versions of the theory, other factors affecting ranking have been considered as well. Rambow (1993) proposed an account of scrambling in German based on the idea that ranking is mainly determined by surface order of realization. The idea that order of mention affects the salience of discourse entities is quite well-established in the psychological literature; experiments based on probe-words lead researchers such as Corbett and Chang (1983); Gernsbacher and Hargreaves (1988)

to 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 - e.g., by Gordon et al. (1993), who observed a REPEATED NAME PENALTY (RNP)[7] for CFs in subject position 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 the first mentioned CF is as highly ranked as the subject.

Strube and Hahn (1999) argue that the rank of discourse entities is determined by the position they hold in Prince's (1981; 1992) givenness hierarchy. More specifically, Strube and Hahn argue that HEARER-OLD entities rank higher than MEDIATED entities; and in turn, these rank higher than HEARER-NEW entities.[8]

$$\text{HEARER-OLD} \prec \text{MEDIATED} \prec \text{HEARER-NEW}.$$

This basic ranking combines with order of mention. Among each category, the entities occurring 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.

Sidner's original hypothesis that ranking depended on thematic roles, abandoned in the early versions of Centering Theory, was put forward again by Cote (1998). This claim 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. 2001b). 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. 2001b,a).

Segmentation and the relationship between global focus and local focus Neither Grosz et al. (1995) nor Grosz and Sidner (1986) give a completely explicit account of the interaction between the two levels of coherence and salience (global and local) assumed in the framework, but subsequent studies have addressed some of the issues raised by this assumption. As far as coherence is concerned, one important question is whether local coherence is completely dependent on global coherence: i.e., whether a shift at the intentional / global level always results in a shift at the local / entity level as well. A number of recent studies and proposals suggest that the relation between the two

[7] Gordon *et al.* introduced the term 'Repeated Name Penalty' to describe the increased reading times they observed when proper names were used instead of pronouns in these conditions, as in *Bruno was the bully of the neighborhood. Bruno / He often taunted Tommy.*

[8] 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.

levels is likely to be more complex. Studies including (Passonneau 1998; Walker 1998) suggest that segment boundaries do not correlate very well with transitions, in that continuations can often 'straddle' segment boundaries, and shifts regularly occur within segments. Work such as (Knott et al. ress) suggests that in genres such as museum descriptions, global coherence may be ensured by relations between entities, whereas local coherence may be of the intentional type.

As far as salience is concerned, a number of studies addressed 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)).

A full investigation of these issues would require a corpus annotated for intentional structure, which is a problem given that identifying segments is still a bit of a black art. We discuss below the heuristics we adopted here; a fuller discussion is in (Poesio and Di Eugenio 2001). We will not be concerned here with 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. ress). For a discussion of these issues, see (Poesio and Di Eugenio 2001).

Different definitions of the central notions of Centering

Alternative definitions of CB Constraint 1 captures Grosz et al's intuition that there is a single 'focus', motivated by the contrasts in acceptability between the discourses above. In the form presented above, the constraint also expresses a strong claim about 'linking' between utterances - namely, that each utterance in a segment realizes at least one of the CFs realized in the previous utterance. A weaker form of the constraint has therefore also been suggested ((Walker et al. 1998a, footnote 2, p.3)):

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

Gordon et al. (1993) propose to replace the definition of CB seen above (Constraint 3) with an operational one: 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 discussed above (a slower reading time whenever a full NP is used instead of a pronoun). Their experiments suggest that RNP effects 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 in subject or first-mention position in the previous utterance.[9] This new definition creates a conflict between this version of Centering Theory and the versions that use Constraint 3 to define the CB; indeed, the

[9]Notice that this claim is different from the claim 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.

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.

An operational definition of the CB was also proposed by Passonneau (1993) on the basis of her study 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.

Alternative Claims about Pronominalization Alternative hypotheses about the relation between Centering and pronominalization have also been advanced. The original formulation of Rule 1 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.

This formulation was subsequently weakened to give the version discussed above. Conversely, Gordon et al. (1993) proposely an even stronger formulation:

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

(Notice that the definition of CB proposed by Gordon *et al.* results in many fewer utterances having a CB.)

One question that, as far as we know, has never been raised is what 'pronouns' exactly should count as pronouns: only third person singular pronouns? What about plural ones, demonstrative pronouns, first and second person pronouns? We will use the term R1-PRONOUN to indicate the (sub) class of pronouns subject to Rule 1.

Competing View of Transitions Rule 2 as formulated by Grosz *et al.* expresses preferences among *sequences* of transitions. Several other versions of this constraint have been proposed, as well as other schemes for classifying transitions. Some of these alternatives 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).

The work just mentioned provides one of the motivations for the formulation of Rule 2 as stating preferences for certain sequences of transitions (e.g., CON-CON over SHIFT-SHIFT) rather than for certain transitions. Di Eugenio (1998), for example, found that the distribution of pronouns depends on the previous transition as well: in continuations that follow a continuation or a shift, it is much more likely that a null pronoun will be used, whereas in continuations that follow a retaining transition, both null and explicit pronouns are equally likely. Turan (1995) found similar results for null and explicit pronouns in Turkish.

Among the researchers arguing that the inferential load is evaluated utterance by utterance, are Brennan et al. (1987), Walker et al. (1994) and (Walker et al. 1998a). Their version of Rule 2 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 introduced by Brennan et al: SMOOTH SHIFT, when $\text{CB}(U_n) = \text{CP}(U_n)$ and ROUGH-SHIFT, when $\text{CB}(U_n) \neq \text{CP}(U_n)$ The result is that transitions can be classified along two dimensions, as in the following table:

	$\text{CB}(U_n) = \text{CB}(U_{n-1})$ or $\text{CB}(U_{n-1}) = \text{NIL}$	$\text{CB}(U_n) \neq \text{CB}(U_{n-1})$
$\text{CB}(U_n) = \text{CP}(U_n)$	CONTINUE	SMOOTH-SHIFT
$\text{CB}(U_n) \neq \text{CP}(U_n)$	RETAIN	ROUGH-SHIFT

Further refinements of the classification scheme for transitions have to do with the classification of utterances that follow an utterance without a CB, such as the first utterance of a segment.[10] Kameyama (1986) proposed a fourth type of transition for these cases, CENTER ESTABLISHMENT; this transition is used by (Di Eugenio 1998) as well. Walker et al. (1994) proposed instead 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.[11]

Strube and Hahn (1999) argue that inferential load should be evaluated across sequences (pairs, in fact) of transitions, but their version of Rule 2 is based on a different way of evaluating the inferential load of utterances. Strube and Hahn argue that other classification of utterances do not reflect what should be one of the crucial claims of the theory - namely, that the CP of an utterance should predict the CB of the next. For this reason, they introduce a distinction between CHEAP and EXPENSIVE transitions (p.332):

- A transition pair is CHEAP if the backward-looking center of the current utterance is correctly predicted by the preferred center of the previous utterance, i.e., if $\text{CB}(U_n) = \text{CP}(U_{n-1})$;
- 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 $\text{CB}(U_n) \neq \text{CP}(U_{n-1})$;

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

[10] According to the strong version of Constraint 1, this is in fact the only utterance that may not have a CB in a coherent discourse.

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

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

Finally, Kibble (pear) 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. His form of Rule 2 is as follows:

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

Salience: prefer transitions such that $\text{CB}(U_n) = \text{CP}(U_n)$.

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

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

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.[12]

Applications

The primary application of theories of text coherence in NLP has been in the development of text planners. Most of the best known such planners are based on relation-centered theories of coherence such as Rhetorical Structures Theory (RST) (Mann and Thompson 1988) (used, e.g., in (Hovy 1993; McKeown 1985; Moore 1995)). However, ideas from the Centering framework are found increasingly useful to supplement a relational notion of coherence (Kibble and Power 2000; Knott et al. ress). The ideas about salience proposed in Centering have been applied to develop algorithms for both anaphora resolution (Brennan et al. 1987; Strube and Hahn 1999; Tetreault 1999) and for sentence planning (Dale 1992; Hitzeman et al. 1997; Henschel et al. 2000).

3 METHODS

We used corpus annotation to compare the different versions of Centering Theory discussed in the previous section and, more in general, to evaluate the claims of the theory (in its 'best variant', if one exists). In this section we discuss how we set about doing this, the data we used, our annotation methods, and how the annotation was used.

[12] See (Karamanis 2001) for further discussion and an evaluation of the effect of these two ways of resolving conflicts.

Verifying the Claims of a Parametric Theory

Quite a few methodological issues have to be considered when trying to evaluate Centering Theory. The first problem is to be clear about what the main claims of the theory are. The development of a 'conceptual vocabulary' for theories of local coherence and local salience is a very significant contribution, but one that is very difficult to evaluate. Instead, we identified Constraint 1, Rule 1, and Rule 2 as main claims of the theory. Even so, we had to take into account the fact that the different versions of Centering Theory sometimes use different definitions of the CB and/or make different claims about coherence and salience; therefore, we considered more than one version of them.

In doing so, we have to be clear about how these claims should be interpreted. The proponents of Centering have been quite explicit that the theory should not be interpreted as stating '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 violated, make a text harder to read, and whose violation has therefore to be signalled in some way. So, the mere presence of a few exceptions to the claims should not count as a falsification. Instead, we will assume that these claims should be verifiable in a statistical sense: the number of utterances that verify such claims should be significantly higher than the number of utterances that violate them. (In most cases, the sign test will be used to test this.)

But 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, 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).[13] 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 configurations 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.

This comparison would be prohibitively expensive with traditional psychological methods, but it's not easy to do with corpus analysis, either. Obviously, it can't be done by directly annotating 'utterances' or 'CB' according to one way of fixing the parameters, as done in most previous studies of Centering Theory (Byron and Stent 1998; Di Eugenio 1998; Kameyama 1998; Passonneau 1993; Walker 1989). Instead, we annotated our corpus with the primitive concepts used by different versions of the theory, i.e., information that has been claimed by one or the other version of Centering to play a role in the definitions of its basic notions. This includes, for example, the grammatical function of an NP, information about anaphoric relations (including information about bridging references) and how sentences break up into clauses and subclausal units. We then used the annotated corpus to compute utterances, their CF ranking, and their CB, according to a particular way of setting the parameters; so that we could then count verifications and violations of the three claims according to

[13] A few studies compare two versions of the theory: e.g., two or more algorithms for anaphora resolution (Strube and Hahn 1999; Tetreault 1999), 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).

that version. We then evaluated each of the claims with respect to a given configuration, and compared the configurations.

A final characteristics of this study is that we were interested in evaluating the claims of the theory in domains of interest for real applications–Natural Language Generation, in our case. The genres most often used to study Centering Theory are 'naturalistic' ones such as narratives or spoken dialogues. This makes a lot of sense from a scientific point of view, but one is left wondering whether the preferences about coherence and salience expressed by Centering Theory might not be overridden by other factors in different genres. For this reason, we used texts in two genres not yet studied in the Centering literature, but of interest to developers of NLG systems: instructional texts and descriptions of museum objects to be displayed on Web pages.

The Data

The data used in this work are texts from the GNOME corpus, that currently includes texts from three domains. The museum subcorpus consists of descriptions of museum objects and brief texts about the artists that produced them.[14] The pharmaceutical subcorpus is a selection of leaflets providing the patients with the legally mandatory information about their medicine.[15] The tutorial dialogues subcorpus consists of a subset of 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; in this study we used texts from the first two domains, for a total of about 3,000 NPs, including 217 personal and possessive pronouns, and 23 demonstratives. As for utterances, the corpus includes about 500 sentences, and 900 finite clauses; the actual number of utterances used in the study is one of the parameters that we varied, as discussed below.

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 fairly specific manual, available from the GNOME project's home page at `http://www.hcrc.ed.ac.uk/~gnome`; in the following we briefly discuss the information that we were able to annotate, what we didn't annotate, and the problems we encountered. Eight paid annotators were involved in the reliability studies and the annotation.

[14] The museum subcorpus extends the corpus collected to support the ILEX and SOLE projects at the University of Edinburgh. ILEX generates Web pages describing museum objects on the basis of the perceived status of its user's knowledge and of the objects she previously looked at (Oberlander et al. 1998); the latest official Web-based demo can be seen at `http://cirrus.dai.ed.ac.uk:8000/ilex`. The SOLE project extended ILEX with concept-to-speech abilities, using linguistic information to control intonation (Hitzeman et al. 1998).

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

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 texts that we thought could be claimed to update the local focus. This includes sentences (defined roughly speaking 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 we felt 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.)[16]

(8)
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))
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 (8d) above illustrates, subordinate units such as clausal complements and relative clauses were enclosed within the superordinate unit. 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 following attributes of units were marked:

- 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 annotator training, but in the end it could be done reliably as well. 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:

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

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

The main problems we 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*:

(9) 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 (10a). However, it wasn't completely obvious what to do in the case of coordination within a subordinate clause, as in (10b). 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 (10c) (where the unit-coordination is marked with square brackets).

(10) 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)

In identifying possible utterances we also to had to address two problems raised by our genres 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 (11).

(11) (u1) Side effects

(u2) 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 (12), 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.

(12) (u1) <unit>Side effects</unit>

(u2) <p> Side effects may occur when PRODUCT-Y is applied to large parts of the body,

Finally, the elements of text that we did *not* mark 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 (as in (13a)), and quoted parts of text, when they are not reported speech (as in (13b)).

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

b. (The inscription 'CHNETOC BASHLHKOC CPATHARHC')

Concerning attributes, one problem we had (especially with the pharmaceutical texts) was instructions in the imperative form, as in (14). 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.

(14) (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).)[17]

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), which in turn were derived from those proposed by Passonneau (1997) and in MUC-7 (Chinchor and Sundheim 1995). We annotated 14 attributes

[17]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.

of NPs specifying their syntactic, semantic and discourse properties (Poesio 2000). Those that are relevant to the work discussed here include:

- The NP type, cat. This attribute can take the values 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*).

- A few other 'basic' syntactic features, num, per, and gen, used to identify contexts in which the antecedent of a pronoun could be identified unambiguously;

- The grammatical function of the NP, gf. Our instructions for this feature are derived from those used in the FRAMENET project ((Baker et al. 1998); see also the project's Web site at `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 (used 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 the relevant 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).

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.*

We encountered all sorts of problems with marking up attributes, even for supposedly 'easy' information such as number and gender, but especially so with semantic attributes (cfr. instructions). Ultimately however we were able to mark up the attributes relevant for this study in a fairly reliable fashion. However, we haven't so far been 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 $\kappa = .35$.

Total number of NPs:	3376
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

Anaphoric information Finally, in order to compute whether a CF in an utterance was realized directly or indirectly in the following utterance, we marked up anaphoric relations between ⟨ne⟩ elements, again using a variant of the MATE scheme (Poesio et al. 1999). A special ⟨ante⟩ element is used to mark anaphoric relations; the ⟨ante⟩ element itself specifies the index of the anaphoric expression and the type of semantic relation (e.g., identity), whereas one or more embedded ⟨anchor⟩ elements indicate possible antecedents (the presence of more than one ⟨anchor⟩ element indicates that the anaphoric expression is ambiguous). (See (15).)

(15)
```
<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 early experiments with Centering, suggested that indirect realization can play quite a crucial role in maintaining the CB. However, previous work, 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, we did annotate this type of relations, but to achieve a reasonable agreement, and to contain somehow the annotators' work, we limited the types of relations annotators were supposed to mark up, and we specified priorities. Thus, besides identity (IDENT) we only marked up three non-identity ('bridging' (Clark 1977)) relations, and only relations between objects (and not, for example, anaphoric reference to propositions or

events). The relations we mark up are a subset of those proposed in the 'extended relations' version of the MATE scheme (Poesio et al. 1999) and include set membership (ELEMENT), subset (SUBSET), and 'generalized possession' (POSS), which includes part-of relations as well as more traditional ownership relations. In addition, given the intended use of this information, we had to specify quite strictly which antecedent should be marked: whereas in MUC it is perfectly acceptable to mark an 'antecedent' which *follows* a given anaphoric expression, in order to compute the CB of an utterance it is necessary to identify the *closest previous* antecedent. Furthermore, we specified preferences concerning NPs occurring in predicative position, so that, for example, in *Francois, the Dauphin*, the embedding NP would be marked as an antecedent, rather than the NP in appositive position.

As expected, we achieved a rather good 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 these relations were marked up by both annotators; 12.8% by only one of them; and in 7.7% of the cases, one of the annotators marked up a closer antecedent than the other. Concerning bridging 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 Although Grosz and Sidner's claims about the global structure of a discourse and its segmentation are not part of Centering Theory per se, the theory does assume that discourses are segmented. This means that, ideally, a corpus used to investigate the claims of the theory should be segmented.[18] The problem is that discourse segments are difficult to identify reliably (Passonneau and Litman 1993; Marcu et al. 1999); our own preliminary experiments didn't give good results, either.

For this reason, most previous studies either ignored segmentation, or used the heuristics proposed by Walker (1989). We did the same here, and only used the layout structure of the texts as a rough indication of discourse structure. We tested both 'looser' forms of segmentation and more fine grained ones based on paragraphs. In the museum domain, the looser segmentation involved treating each object description as a separate segment; in the pharmaceutical domain, each subsection of a leaflet was treated as a separate segment. The finer-grained segmentation was the one proposed by Walker. We then identified by hand those violations of Constraint 1 that appeared to be motivated by too broad a segmentation of the text.[19]

Automatic computation of Centering information

The annotation thus produced was used to automatically compute utterances according to the particular configuration of parameters chosen, 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. This information was the used to find violations of Constraint 1, Rule 1, and Rule 2 (according to

[18] This has been contested in recent 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)).

[19] Work such as (Moser and Moore 1996b; Marcu et al. 1999) showed that it is indeed possible to achieve good agreement on discourse segmentation, given intensive training and repeated iterations. We are making use of a corpus reliably annotated in this way in other work (Poesio and Di Eugenio 2001).

several versions of Rule 1 and Rule 2). The behavior of the script that computes this information depends on the following parameters:

CBdef : whether Grosz Joshi and Weinstein's, Gordon et al's, or Passonneau's definition of CB should be used.

uttdef: whether utterances should be identified with sentences, finite clauses, or verbed clauses.

previous utterance: whether adjunct clauses should be treated Kameyama-style or Suri-style.[20]

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

realization: whether only direct realization should be allowed, or also indirect realization via bridging references.

CF-filter: whether all NPs should be treated as introducing CFs, or whether certain classes should be excluded (currently the possible omissions include second and first person NPs and NPs in predicative position (e.g., *a policeman* in *John is a policeman*).

rank: whether CFs should be ranked according to grammatical function, linear order, a combination of the two as suggested by Gordon *et al.*, or information status in Strube and Hahn's sense.

prodef: whether only third person personal pronouns like *it, they* should be counted as pronouns for the purposes of Rule 1, or also demonstrative pronouns like *that, these* and / or the second person pronoun *you*.

segmentation: 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.[21]

Rule 1

The way the various statistics reported below are computed is mostly transparent; the only aspect that needs discussing are the computations for Rule 1. The basic logic is very simple: for each utterance *u*

1. If *u* has no CB, it is ignored;

[20]In fact, the version we call here 'Kameyama' treats *all* types of clauses other than complement clauses – including, e.g., relative clauses–as not embedded, whereas the version 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' version assigns the relative clause *which is one of the refractory metals* as previous utterance to the clause *It was made by ...*; whereas the 'Suri' version treats the relative clause as embedded. We return to this issue below.

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

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,

   (a) Count *u* as a violation (-) of Gordon *et al.*'s version of Rule 1;

   (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), else as a +.

The one additional complication are relative pronouns. As their status for the theory is not clear, we decided to ignore them as much as possible, in the following sense: the script does not count an utterance as a violation of Rule 1 from (Grosz et al. 1995) if the only 'pronoun' realizing a non-CB is a relative pronoun; and conversely, it does not count an utterance as a verification of that Rule if the CB is only realized by a relative pronoun. The main consequence is that the number of utterances taken into account for Rule 1 is generally less than the number of utterances with a CB, as we will see shortly.

4 MAIN RESULTS

Given that there are so many parameters, it is difficult, if not impossible, to evaluate all versions of the theory. Instead, we began by identifying a 'vanilla configuration' of the theory based on the most familiar choices about the parameters, and we tested the claims of the theory given these values. We then studied the versions obtained by varying the 'minor' parameters: utterance, realization, and segmentation.[22] After establishing the 'best' values for these parameters, we looked at the effect on the claims of alternative ranking functions, and finally we varied the definition of CB.

The Vanilla Configuration

What we call 'vanilla' version here is a blend of proposals from (Grosz et al. 1995) and (Brennan et al. 1987), with additional suggestions from (Kameyama 1998), (Gordon et al. 1999), and (Walker 1998). It is based on the definition of CB from (Grosz et al. 1995), and, for ranking, on the proposal that CFs are ranked according to grammatical function, as discussed there and in (Brennan et al. 1987) (also incorporating the proposals concerning ranking in complex NPs from (Gordon et al. 1999)). As far as utterance definition is concerned, the vanilla version incorporates the hypothesis from Kameyama (1998) that utterances are finite clauses, and the characterization of 'previous utterance' proposed there;[23]. Concerning realization, only third person NPs are taken to introduce CFs (not first or second person); and a discourse entity only counts as 'realized' in an utterance if it is explicitly mentioned. For the purposes of Rule 1, we consider both a 'strict' definition of 'pronoun' including only personal and possessive pronouns, and a 'broad' one including also the demonstrative pronouns *this*, *that*, *these* and *those*. As for relative clauses, we assume that they include a link to the embedding NP, possibly

[22] We refer to these parameters as 'minor' because they haven't been the focus of as main research as ranking and the definition of CB.

[23] We simplified Kameyama's hypothesis about relative clauses by comparing only a version in which they were treated as utterances both 'locally' and 'globally', and one in which they weren't.

not explicitly realized.[24] For segmentation, we adopt the segmentation heuristic proposed by Walker (1989). With the parameters of the theory defined this way, definitions, the number of utterances and CFs in our corpus is as follows:

	MUSEUM	PHARMA	TOTAL
Number of utterances:	428	577	1005
Number of CFs:	1723	1308	3031

Constraint 1 The statistics relevant to Constraint 1 (that utterances have exactly one / at most one CB) are shown in the following table:

	MUSEUM	PHARMA	TOTAL (PERC)
Number of times at least one CF(Un) is realized in Un+1:	197	165	362 (36%)
Utterances that satisfy Constraint 1 (have exactly one CB) :	193	160	353 (35.12%)
Utterances that do not satisfy C1 but are segment boundary:	66	96	162 (16.11%)
Utterances with zero CBs :	165	316	481 (47.86%)
Utterances with more than one CB :	4	5	9 (0.8%)

These figures clearly indicate that the weak version of Constraint 1, verified by 834 utterances (82.98%) and violated by 9 (.8%) (abbreviated henceforth as +834, -9) is likely to hold with the 'Vanilla' version of the theory (a sign test indicated a chance $p \leq 0.001$ that Weak C1 does not hold with other samples). On the other hand, the strong version of C1 –that every utterance has exactly one CB)–is not likely to hold: in our corpus, only 353 utterances out of 1005 (35.12%) have exactly one CB, and even if we exclude the 162 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, there are still 490 utterances with zero or more than one CB (48.75%). With +353, -490, a sign test indicates that the chance of error in rejecting the null hypothesis that Strong C1 doesn't hold is much higher than 10%.[25]

The following example illustrates one class of counterexamples to Strong C1 with the Vanilla setting. In (16), if we identify utterances with finite clauses, u1 is followed by four utterances. Only the last of these directly refers to the set of egg vases introduced in u1, while they all contain implicit references to these objects. In (16a), (entity) coherence is maintained by the bridging reference (*the furniture*) rather than by direct reference. Clearly, there are two ways of 'fixing' this problem with the Vanilla version: either identifying utterances with sentences, in which case utterances (u2)-(u4) disappear; or allowing for indirect realization, in which case (u2)-(u4) all will have a CB. We will consider both of these possibilities below.

(16) (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: ...

Perhaps even more interesting is the fact that several utterances have *more than one* CB - i.e., they violate Weak C1 as well. This is illustrated by (17), where we kept the XML format of the annotation so as the attributes of elements were included.

(17)
```
<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>
```

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

[25] Furthermore, the figure of 353 utterances verifying Strong C1 includes 71 relative clauses whose only reference to entities in the embedding clause is their complementizer or a trace.

```
   <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>
```

In this example, two discourse entities introduced in utterance u227 are realized in utterance u229:[26] *the corner cupboard* (realized by ne547 and ne549) and *Branicki* (realized by ne551). 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 for examples), 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.

Salience and Pronominalization The statistics concerning pronominalization and the CB are shown in the following table. R1 pronouns include personal pronouns and relative pronouns / traces; the figures concerning demonstrative pronouns are also listed.

	MUSEUM	PHARMA	TOTAL
Total number of R1-pronouns:	271	120	391
Number of personal pronouns:	144	73	217
Number of relative pronouns:	127	47	174
Number of demonstrative pronouns:	7	16	23
Utterances with a subject:	383	216	599
Number of personal pronouns in subject position:	61	34	95
Number of demonstrative pronouns in subject position:	5	11	16
Total number of realizations of CBs:	218	166	384
Total number of CBs realized as R1-pronouns:	144	69	213
CBs realized as personal pronouns:	91	49	140
CBs realized as relative pronouns:	53	20	73
CBs realized as demonstrative pronouns:	3	1	4
CBs NOT realized as R1-pronouns:	74	97	171
Total number of R1-pronouns that do not realize CBs:	53	22	75
Personal pronouns that do not realize CBs:	51	20	71
Relative pronouns that do not realize CBs:	2	2	4
Demonstrative pronouns that do not realize CBs:	4	15	19

Our corpus includes 217 uses of personal pronouns (*he*, *she*, *it*, *they*, and their other morphological forms), 23 demonstratives, and 174 relative pronouns or traces, for a total of 391 R1-pronouns

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

(counting relative pronouns or traces). Of the personal pronouns, 37 (17%) have their antecedent in the same utterance, and 28 (13%) in an utterance further away than the previous utterance. The corpus contains 59 pronoun-pronoun chains (cases in which the antecedent of a pronoun is itself realized as a pronoun). Of the 353 utterances with exactly one CB, 72 are ignored by the script in that the only realization of a R1-pronoun is done via a relative pronoun or trace, and 281 are considered as relevant for Rule 1.[27]

The first thing to notice is that unless we count relative pronouns and relative traces as R1-pronouns, about as many - in fact, more - CBs are realized as non-pronouns than as pronouns (140 CBs are realized as personal pronouns; 73 as relative pronouns or traces; and 171 as non-pronouns). What this means is that the stronger version of Rule 1 proposed by Gordon et al. (1993) (always pronominalize the CB) only holds (and then with a 7% chance of error) if we count relative pronouns as R1-pronouns (see also (Henschel et al. 2000)). On the other hand, both the version of Rule 1 originally proposed by (Grosz et al. 1983) and that in (Grosz et al. 1995) do hold. The complete figures concerning satisfaction and violation of the three versions of Rule 1 discussed in Section §2 are shown in the following table.[28]

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	135	138	273 (97.1%)
GJW 95 - utterances that violate:	6	2	8 (2.8%)
Gordon - utterances that satisfy:	81	46	127 (45.2%)
Gordon - utterances that violate:	60	94	154 (54.8%)
GJW 83 - utterances that satisfy:	122	109	231 (82.2%)
GJW 83 - utterances that violate:	19	31	50 (17.8%)

Two examples of utterances violating Grosz *et al.*'s version of Rule 1, which requires the CB to be pronominalized if anything else is, are shown in (18).

(18) a. (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.

b. (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 (18a), the CB of u3 is *Louis XIV*, the king, which is however realized using a proper name, presumably because of the reference to an official title; the pronoun *he* is used to realize *Boulle*, which,

[27] See the description of the way the statistics for Rule 1 are computed in Section §3.

[28] We are using here the 'narrow' definition of R1-pronoun, which only includes personal pronouns, and not demonstrative and relative pronouns.

while the 'main character' in the sense of Garrod and Sanford of this discourse (and the 'discourse focus' in the sense of Sidner), is not the CB of u3. In other words, we can observe here a conflict between the idea that pronominalization is used to realize the 'main entity' of a discourse, irrespective of its ranking, and the idea that pronominalization is used to realize the locally most salient entity, as identified by the CB. (See also (Giouli 1996; Byron and Stent 1998).) In (18b), the CB of u3 is PRODUCT-X, which, however, is realized using a proper noun, whereas a possessive pronoun is used to refer intrasententially to *the baby* (For a discussion of the problem of intrasentential pronouns in Centering Theory, see (Walker 1989; Tetreault 1999; Poesio and Stevenson pear)).

In the pharmaceutical leaflets we found a number of violations of Rule 1 towards the end of texts, when a number of pronouns are used to realize the product described by the leaflet. E.g., *it* in the following example refers to the cream, not mentioned in any of the previous two utterances.

(19) (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.

These cases may be seen again as examples of the conflict between the 'global' preference to realize the 'main character' and the 'local' preference to realize the most highly ranked entity. By the end of the text, after the product has been mentioned a number of times, it is salient enough that there is no need to put it again in the local focus by mentioning it explicitly.

The results change only slightly when a 'wider' sense of pronoun is adopted by considering demonstrative pronouns as well (but see (Passonneau 1993)): in this case, we have more violations of the version of Rule 1 from (Grosz et al. 1995) (10 instead of 8) but fewer violations of the version of Gordon *et al.* (150 instead of 154) and of the version in (Grosz et al. 1983) (48 instead of 50). (In the rest of the paper we will keep using the 'narrow' definition of pronoun.)

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	135	136	271
GJW 95 - utterances that violate:	6	4	10
Gordon - utterances that satisfy:	84	47	131
Gordon - utterances that violate:	57	93	150
GJW 83 - utterances that satisfy:	124	109	233
GJW 83 - utterances that violate:	17	31	48

One interesting effect of this change is that if we count demonstrative pronouns among the pronouns governed by Rule 1, we get more long-distance pronouns (39) than intra-utterance ones (37).

Rule 2 The figures concerning transitions relevant for Brennan et al's version of Rule 2 are shown in the following table.

	MUSEUM	PHARMA	TOTAL
Establishment :	96	95	191 (19%)
Continuation :	37	32	69 (6.8%)
Retain :	24	17	41 (4%)
Smooth Shift :	19	13	32 (3.2%)
Rough Shift :	21	8	29 (2.9%)
Zero :	86	81	167 (16.7%)
Null :	145	331	476 (47.4%)
Total :	428	577	1005

The most interesting result here is that the most frequent transition by far, with 47% of the total, is one not mentioned in the Centering literature, the NULL transition, which connects two utterances without a CB. The second most common transition 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 from an utterance with a CB to one without (also not mentioned in the literature), and then by CON, RET, SSH, and RSH. If we ignore NULL transitions and ZEROs, the preferences are roughly as predicted[29] by Brennan *et al.*, especially if we merge EST with CON as suggested by Walker et al. (1994); there are about the same number of RSH and SSH. (Similar results were obtained by (Passonneau 1998).) 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 really useful:

	MUSEUM	PHARMA	TOTAL
Continuations Sequences :	10	5	15
Establishment /Continuation :	16	18	34
Retain Sequences :	6	3	9
Retain / Smooth Shift :	3	1	4
Retain / Rough Shift :	3	2	5
Smooth Shift Sequences :	2	1	3
Rough Shift Sequences :	3	1	4
Zero Sequences :	0	0	0
Null Sequences :	90	229	319
Other :	227	280	507

(We should add that we used the most favourable way of counting sequences–each pair of repeated transitions was counted as a sequence, which means that three CONT in a row count as two sequences.) In our corpus there seems to be a preference for avoiding repetition, even in the type of transitions: e.g., EST / CONT sequences are twice as common as sequences of continuations.

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 version, which is not the one assumed by Strube and Hahn themselves (but see below); this confirms results obtained for dialogues by Byron and Stent (1998). Ignoring the 225 utterances which are segment boundaries,[30] we have 401 pairs of expensive transitions, and 32 pairs of cheap transitions, as follows:

	MUSEUM	PHARMA	TOTAL
Cheap transitions :	76	65	141
Expensive transitions :	261	378	639
Cheap transition pairs :	18	14	32
Expensive transition pairs :	161	240	401

Finally, we devised the following method to evaluate Kibble's proposal. 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 configuration, the average Kibble score[31] comes to about 1.06 - i.e., each utterance satisfies about one of the four constraints. The figures are as follows:

[29] Unfortunately we are not aware of a test that could be used to verify a claim about the ranking of relative frequencies, like Rule 2.

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

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

	MUSEUM	PHARMA	TOTAL
Continuous transitions :	197	165	362
Salient transitions :	105	108	213
Cheap transitions :	150	128	278
Cohesive transitions :	60	51	111
Average 'Kibble Score' :	1.52	1.02	1.24

Differences between domains: Broadly speaking, 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. Counting personal pronouns only, there are fewer pronouns in the pharmaceutical domain (73 of 1308 CFs, or 5%, as opposed to 144 of 1723, 8%, for the museum domain), and whereas in the museum domain 41.7% (91/218) of CB realizations are done via personal pronouns (66% if we also count relative pronouns and complementizers), in the pharmaceutical domain only 29.5% (49/166) are (41% with relative pronouns). The percentage of utterances satisfying the strong version of Constraint 1 is much higher in the museum domain (45%, 193/428) than in the pharmaceutical domain (27.7%, 160/577), and the percentage of utterances with no CB is much higher in this second domain (54%, 316/577) than in the first one (38%, 165/428). Finally, over 71% of utterances in the pharmaceutical domain are NULL or ZERO transitions (412/577), whereas just 53% are in the museum domain (231/428); the percentage of EST and CONT is also slightly higher in the museum domain (133 / 428, 31%, versus 127 / 577, 22%).

As discussed below, these differences are in part be due to the large number of second person pronouns *you* in the pharmaceutical domain, many of which serve to maintain coherence and / or as most salient entities.

Varying the utterance parameters

In this subsection we consider how changing the definition of utterance and of previous utterance affects Constraint 1, Rule 1 and Rule 2.

Treating coordinated VPs as utterances Several researchers studying spoken dialogues have suggested that each element of a coordinated VP should be treated 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 would be especially natural in grammatical theories in which coordinated VPs are viewed as sentences with an empty subject. In the texts in our corpus, however, treating coordinated VPs as separate utterances leads to slightly worse results, mainly because more units count as utterances (1039 vs. 1005 with the Vanilla version). The differences are significant for Constraint 1 (30 additional violations) but not for Rule 1.[32] The relevant figures for Constraint 1 are as follows:

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1:	201	165	366
Utterances that satisfy Constraint 1 (have exactly one CB) :	197	160	357
Utterances that do not satisfy Con 1 but are segment boundary:	66	96	162
Utterances with zero CBs :	179	332	511
Utterances with more than one CB :	4	5	9

[32]We should however remark that we didn't treat these coordinated VPs as containing a trace; doing so might lead to better results.

whereas those for Rule 1 (counting relative pronouns as R1-pronouns) are:

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	136	139	275
GJW 95 - utterances that violate:	8	1	9
Gordon - utterances that satisfy:	83	46	129
Gordon - utterances that violate:	61	94	155
GJW 83 - utterances that satisfy:	124	109	233
GJW 83 - utterances that violate:	20	31	51

The other significant change is in the number of cheap and expensive transitions: treating coordinated VPs as separate utterances results in many more utterances being classified as expensive (36).

	MUSEUM	PHARMA	TOTAL
Cheap transitions :	73	65	138
Expensive transitions :	282	394	676
Cheap transition pairs :	16	13	29
Expensive transition pairs :	180	253	433

Using all verbed clauses instead of just the finite ones A second possible extension of the definition of utterance is to treat *all* clauses with a verb as utterances, rather than just those with a finite verb. However, with this definition we have many more utterances (1266 instead of 1005) and significantly more violations of the strong version of Constraint 1 (685 vs. 490). There are no significant differences in the number of violations of Rule 1. As for Rule 2, this change results in many more NULL transitions and in more EST, about the same number of shifts, and fewer CON and RET, as shown by the following table:

	MUSEUM	PHARMA	TOTAL
Establishments:	141	102	243
Continuations :	30	28	58
Retain :	22	12	34
Smooth Shift :	31	12	43
Rough Shift :	22	8	30
Zero :	123	88	211
Null :	209	438	647

Treating titles and other layout elements as utterances Our 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 truly considering only 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 only in the case of Strong C1 is the difference significant. This is even more true of the case discussed below when utterances are identified with sentences. Titles are treated as utterances in the configurations studied in the rest of the paper, even when they are not finite clauses or sentences.

Other changes to the definition of utterance In general, the only case in which adding more units results in fewer violations of Constraint 1 and Rule 1 is with titles. Otherwise, the best results (especially for C1) are obtained by considering larger text constituents as utterances, thus reducing the number of utterances. In particular, improvements are obtained by eliminating finite clauses that occur as parentheticals, as subjects (as in *That John could do this to Mary was a big surprise to me*), and as matrix clauses with an empty subject (as in *It is likely that John will arrive tomorrow*). This merging of clauses only reduces the overall number of utterances from 1005 to 971, but the result is a simultaneous reduction in the number of violations of Strong C1, from 490 to 464 (which is significant, while still not enough for Strong C1 to be verified by the binomial proportions test) and a small increase in the number of utterances that satisfy Rule 1 (in the version from (Grosz et al. 1995)) to 279, while also reducing the violations to 7 (not significant). There are virtually no changes as far as Rule 2 is concerned. Because of these small improvements, in what follows when we discuss the results with finite clauses as utterances we always exclude these types of finite clauses.

Relative Clauses Finding out the best treatment of relative clauses turned out to be difficult. The reader may recall that Kameyama tentatively proposes (without empirical verification) that relative clauses have a 'mixed' status: they should be locally treated as updating the local focus, but at the global level they should be merged with the embedding utterance. This proposal however seems to involve a final step in which the local focus is updated with the content of certain utterances some time after they have been first processed, which a rather radical change to the basic assumptions of the framework. Instead, we simply considered a version of the theory in which relative clauses are not treated as utterances, and compared it with the versions discussed so far, in which they are. In addition, we compared treating relative clauses as adjuncts (i.e., as not embedded) and treating them as complements (embedded).[33] The figures reported so far were obtained by treating relative clauses as utterances, and as akin to adjuncts; in addition, we have been assuming that relative clauses contain a null element / trace referring to the entity modified by the relative, so that relative clauses never violate C1. This turns out to be the worse configuration. Not treating relative clauses as utterances results in 6% fewer utterances (907 instead of 971) which in turns means significantly fewer violations of Weak C1, 447 (436 utterances without a CB, 11 with two CBs) instead of 464 (454 and 10). The number of violations of Rule 1 stays the same, 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% (from 191 to 158); we also see a 30% reduction in SSH. Everything else stays the same.

In purely numerical terms, then, one could argue that not treating relative clauses as utterances would result in a small improvement. On the other hand, we feel that excluding finite relative clauses would make it very difficult to maintain the principle that utterances are identified with finite clauses. And anyway, we will see in a moment that the additional violations of Constraint 1 also disappear if we treat relative clauses as complements rather than adjuncts, i.e., if we adopt a 'generalized Suri' notion of previous utterance rather than a 'generalized Kameyama'. For these reasons, in the runs discussed in the rest of the paper we continued to treat relative clauses as separate utterances.

Suri and McCoy's definition of previous utterance As discussed in Section Section §2, the experiments of Suri and McCoy suggested that adjunct clauses such as *after* and *before* clauses behaved more like embedding elements (i.e., like complements) than like coordinating ones; Cooreman and Sanford found evidence supporting this treatment for *when* clauses, as well. We tested a version

[33] 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 ... .*

of Centering in which Suri and McCoy's treatment is adopted for all adjuncts; in this version, for example, the previous utterance for (20c) is (20a), whereas in Kameyama's version, it is (20b). We call this version *generalized Suri-McCoy*.

(20) a. *John woke up*
b. *when Bill rang the door.*
c. *He had forgotten the appointment*

Using Suri's definition of previous utterance for embedded adjunct clauses, rather than Kameyama's, results in small but significant improvements concerning Strong C1, as well as in improvements concerning R1, and in no worse results for Rule 2. First of all we have a significant reduction in the number of violations of Constraint 1, although not in all cases is there an improvement: 25 utterances that violate Strong C1 under Kameyama's definition satisfy it under Suri's, but 13 utterances become violations (by the sign test, +25, -13, $p \leq .03$). This reduction is still not sufficient for Strong C1 to be verified.

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1:	198	170	368
Utterances that satisfy Constraint 1 (have exactly one CB):	193	164	357
Utterances that do not satisfy Con 1 but are segment boundary:	67	92	159
Utterances with zero CBs :	139	305	444
Utterances with more than one CB :	5	6	11

The overall figures for 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:[34]

Principle	Plus	Minus	p
CONSTRAINT 1 (STRONG)	357	455	$p = 1.000$
CONSTRAINT 1 (WEAK)	801	11	$p = 0.000$
RULE 1 (GJW 95)	290	7	$p = 0.000$
RULE 1 (GORDON)	135	162	$p = 0.934$
RULE 1 (GJW 83)	246	51	$p = 0.000$

It should be noted, however, that these differences have mostly to do with the way relative clauses are handled, i.e., with examples like the following.

(21) *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.*

If what we call here 'generalized Kameyama' definition of previous utterance is adopted, the previous utterance for the clause *It was made by ...* is the relative clause *which is one of the refractory metals*; this causes causing a violation of Strong C1. The 'Suri' version, by contrast, the relative clause is treated as embedded. If we didn't treat relative clauses as utterances, we would have an equal number of violations for the two versions, although about 20 of these violations would be specific to each version. One example where the difference doesn't have to do with relative clauses, but with the treatment of adjuncts, is (22). PRODUCT-Z is not mentioned in the adjunct *if*-clause, and therefore a violation of Strong C1 results if (u2) is taken as previous utterance for (u3). In this case, Suri and McCoy's treatment of adjuncts leads to a better result than Kameyama's.

[34]The reader may have noticed that 297 utterances are considered for the evaluation of Rule 1, rather than 281 for the Vanilla version. This is because of the way the algorithm for counting violations of Rule 1 works: different numbers of utterances may be considered depending on the presence of relative pronouns.

(22) (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 hypothesis gives the right result.

(23) (u1) Do not use scissors

(u2) as you may damage the patch inside.

(u3) Take out the patch.

Suri's definition of previous utterance –more precisely, treating relative clauses and all types of adjuncts as embedded –also leads to better results concerning Rule 2: fewer NULL and ZERO transitions, more Center Establishments and Center Continuations, more SSH than RSH, more cheap transitions, fewer expensive ones, and a better 'Kibble Score' (1.14 instead of 1.09). The differences between the 'generalized Suri' version 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.

Sentences By far the most dramatic improvement as far as Strong C1 is concerned result from identifying utterances with sentences; in fact, the improvement is such that under certain conditions Strong C1 becomes verified. If we only count sentences as utterances, the number of utterances goes down quite considerably, by almost 50% (from 1005 to 535), and the number of utterances with zero CBs also halves. However, if we solely consider sentences a number of CFs would not belong to any utterance, since many CFs are introduced in titles and other layout elements which do not have a sentential format, such as *Chandelier* or *Side effects*. Just as we did in the case of finite clauses, then, we treat such text constituents as utterances, as well; this brings the total number of utterances to 668. The figures relevant to Constraint 1 with this definition of utterance are:

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1:	130	147	277
Utterances that satisfy Constraint 1 (have exactly one CB) :	126	138	262
Utterances that do not satisfy Con 1 but are segment boundary:	65	80	145
Utterances with zero CBs :	75	173	248
Utterances with more than one CB :	4	9	13

If we only consider the 535 sentences, both Strong and Weak C1 are now verified (the sign test gives $p \leq 0.001$ for Strong C1). However, Strong C1 is not verified if we consider all 668 segments of text that contain NPs: in this case, the number of utterances that satisfy Strong C1, (264) is almost identical with the number of those that don't (261).[35]

Identifying utterances with sentences also has several negative (if small) effects, however. The first of these is that the number of utterances with more than one CB increases in this version by 50% (from 9 in the Vanilla version to 13). This is because many sentences include more than one clause, which increases the likelihood that more than one discourse entity will be realized in the same grammatical function or an equivalent one (remember that the ranking function adopted in the 'vanilla' version of Centering does not include any provision for 'tie-breakers' such as linear order). An example of

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

multi-clausal sentence in which more than one entity is realized in the same grammatical function is the following discourse, where both *the famous Parisian palace, ...* and *the King's cousin, ...* occur in 'OTHER' position in (s73) (in different clauses) and are subsequently mentioned in (s74), which makes both of them potential CBs:

(24) (s73) These four wall lights are among eight made in 1756 for the newly redecorated interiors of the famous Parisian palace, the Palais-Royal, which was the residence of the king's cousin, Louis-Philippe, duc d'Orleans.

(s74) Shortly after inheriting the building in 1752, he commissioned the architect Pierre Contant d'Ivry to renovate the main rooms.

Identifying utterances with sentences also has a a negative effect on Rule 1: again, the number of violations goes up by 50%, from 8 to 12. Because the number of violations is still quite small, both the version of Rule 1 in (Grosz et al. 1995) and the original one in (Grosz et al. 1983) are still verified (+252, -12; and +209, -55, respectively, as opposed to +273, -8 and +231, 50 with the Vanilla version[36]), although Gordon et al's version still isn't (+97, -167). The overall statistics about pronominalization for the version identifying utterances with sentences are as follows:

	MUSEUM	PHARMA	TOTAL
Utterances with a subject:	245	172	417
Total number of R1-pronouns in subject position:	61	34	95
Number of personal pronouns in subject position:	61	34	95
Number of demonstrative pronouns in subject position:	5	11	16
Total number of realizations of CBs:	183	158	341
Total number of CBs realized as R1-pronouns:	89	41	130
CBs realized as personal pronouns:	89	41	130
CBs realized as relative pronouns:	0	0	0
CBs realized as demonstrative pronouns:	4	2	6
CBs NOT realized as R1-pronouns:	94	117	211
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:	107	103	209 (79.2%)
GJW 83 - utterances that violate:	19	36	55 (20.8%)

The results for Rule 2 depend again on whether we only count 'pure' sentences, or all segments of text that contain a CF. With a 'pure' notion of sentence, the number of NULL transitions is drastically reduced (to 162), and the number of both SSH and RSH increases. (In this version the total number of shifts is greater than the number of RET, and even than the number of 'pure' CON.) The figures are as follows:

[36]The reader should keep in mind that the number of utterances to be tested is different depending on whether utterances are identified with finite clauses (281) or sentences (264).

	MUSEUM	PHARMA	TOTAL
Establishments :	49	52	101
Continuations :	26	32	58
Retain :	25	31	56
Smooth Shift :	10	20	30
Rough Shift :	22	19	41
Zero :	44	43	87
Null :	65	97	162

If we also include layout elements where necessary, the results are more similar to those obtained with finite clauses, as follows:

	MUSEUM	PHARMA	TOTAL
Establishments :	54	68	122 (18.3%)
Continuations :	28	33	61 (9.1%)
Retain :	22	23	45 (6.7%)
Smooth Shift :	7	12	19 (2.8%)
Rough Shift :	19	11	30 (4.5%)
Zero :	52	66	118 (16.7%)
Null :	88	185	273 (40.9%)

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).

The figures for the sentences-only version are as follows:

	MUSEUM	PHARMA	TOTAL
Continuations Sequences :	10	9	19
Establishment /Continuation :	11	14	25
Retain Sequences :	4	5	9
Retain / Smooth Shift :	1	2	3
Retain / Rough Shift :	6	1	7
Smooth Shift Sequences :	0	1	1
Rough Shift Sequences :	4	1	5
Zero Sequences :	0	0	0
Null Sequences :	50	136	186
Other :	176	226	402

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 are more expensive-expensive sequences than cheap-cheap ones.

	MUSEUM	PHARMA	TOTAL
Cheap transitions :	54	44	98
Expensive transitions :	125	220	345
Cheap transition pairs :	11	7	18
Expensive transition pairs :	57	133	190

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

	MUSEUM	PHARMA	TOTAL
Continuous transitions :	130	147	277
Salient transitions :	53	87	140
Cheap transitions :	54	44	98
Cohesive transitions :	50	56	106
Average 'Kibble Score' :	1.60	1.27	1.4

Although the figures just discussed indicate that identifying utterances with sentences leads to better results in many respects, we believe the case is not completely settled. This is in part because of theoretical reasons: e.g., in other theories of discourse where 'units' are assumed, such as RST, these units are generally finite clauses. Secondly, identifying utterances with sentences leads to small, but significant increases in the number of violations of Rule 1 (from 8 in the Vanilla version, 2.8%, to 12, 4.5%) and in the number of Rough Shifts (from 2.9% to 4.5%). But most important of all, we will see in a moment that there are other ways of changing the Vanilla version that also satisfy Strong C1 without identifying utterances with finite clauses, so adopting this definition of utterances is not strictly necessary. For this reason in the rest of the paper we will not simply identify utterances with sentences, but we will also study the effect of the changes to the other parameters on the version in which utterances are identified with finite clauses. For brevity, we will indicate the versions in which utterances are identified with finite clauses as u=f, and the versions in which they are identified with sentences as u=s.

Realization

In this section we discuss the effect of changes on the value of the realization parameter.

IF: Indirect realization + u=f Examples such as (16a) indicate that another way to reduce the number of violations of Constraint 1 is to allow for indirect realization. And indeed, if we modify the 'best' among the u=f versions –that using our generalization of Suri and McCoy's proposals about previous utterances, and which does not count coordinated VPs and parentheticals–to allow for indirect realization, we get a significant improvement for Constraint 1; so much so that even the strong version of the constraint is verified by the sign test (+525, -324). The complete figures for this version are as follows.

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1 :	298	248	546
Utterances that satisfy Constraint 1 (have exactly one CB) :	290	235	525
Utterances that do not satisfy Con 1 but are segment boundary:	48	74	122
Utterances that do not satisfy Con 1 but are relative clauses:	0	0	0
Utterances with zero CBs :	58	245	303
Utterances with more than one CB :	8	13	21

However, allowing for indirect realization has the same two negative effects as the change to u=s. The first is that the number of utterances with more than one CB doubles, from 11 to 21; but because the number of such violations is still relatively small, Strong C1 is still verified. We also find a significant increase in the number of violations of Rule 1, which also double: from 7 with the Suri setting to 14. But because more utterances have a CB with indirect realization, the number of utterances that matter for the purposes of Rule 1 also increases from 281 to 467, so the relative percentages do not change much with respect to the configuration with direct realization (e.g., now 3% of utterances violate Rule 1 in the GJW 95 version, as opposed to 2.3% with generalized Suri and direct realization). Both the

version of Rule 1 –from Grosz et al. (1995) and Brennan *et al.*; and from (Grosz et al. 1983)–are still verified, but not the one by Gordon *et al.*. The overall statistics for pronominalization with this version and u=f are shown in the following table.

	MUSEUM	PHARMA	TOTAL
Total number of realizations of CBs:	225	174	399
Total number of CBs realized as R1-pronouns:	138	74	212
CBs realized as personal pronouns:	98	55	153
CBs realized as relative pronouns:	40	19	59
CBs realized as demonstrative pronouns:	3	1	4
CBs NOT realized as R1-pronouns:	87	100	187
Total number of R1-pronouns that do not realize CBs:	44	15	59
Personal pronouns that do not realize CBs:	41	13	54
Relative pronouns that do not realize CBs:	3	2	5
Demonstrative pronouns that do not realize CBs:	4	15	19

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

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	239	214	453 (97%)
GJW 95 - utterances that violate:	12	2	14 (3%)
Gordon - utterances that satisfy:	87	52	139 (29.8%)
Gordon - utterances that violate:	164	164	328 (70.2%)
GJW 83 - utterances that satisfy:	172	146	318 (68.1%)
GJW 83 - utterances that violate:	79	70	149 (31.9%)

An example of pronominalization that becomes a violation of Rule 1 if we allow for CBs to be indirectly realized is shown in (25). The NP *One stand* in u42 is a bridging reference to the CF 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. (Of course, this pronoun would not count as a violation if the non-finite clause containing *it* were counted as a separate utterance - we saw above however that this move leads to worse results in general.)

(25) (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*.

The change to indirect realization also has an impact on the statistics for transitions. Because these indirect realizations do not occupy the most salient grammatical functions in the new utterance, adopting indirect realization leads to a large increase in the number of retaining transitions. The number of rough shifts greatly increases, as well.

	MUSEUM	PHARMA	TOTAL
Establishments:	74	95	169
Continuations :	50	39	89
Retain :	78	52	130
Smooth Shift :	35	23	58
Rough Shift :	61	39	100
Zero :	59	78	137
Null :	47	241	288

Finally, we find an improvement in the other versions of Rule 2: the percentage of cheap transitions increases (from 153 / 971, 15.7%, to 205 / 971, 21.1%, as opposed to 14.7% for the u=s version) and the Kibble score increases as well, from 1.14 to 1.6 (vs. 1.4 for the u=s version).

In what follows, we will indicate the instantiations of the theory with u=f (and Suri-style treatment of adjuncts) and direct realization as DF; those based on indirect realization as IF.

IS: Indirect realization + u=s As one might expect, even better results for Constraint 1 are obtained by combining indirect realization with the u=s version. With this configuration (henceforth, IS) 389 utterances (out of 668) satisfy the strong version of C1, and 176 violate it; this is significantly better than the u=s version with direct realization (henceforth, DS). Note however that the number of utterances with more than one CB doubles again with respect to the DS version, to 25 (3.7%).

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1:	192	222	414 (62%)
Utterances that satisfy Constraint 1 (have exactly one CB) :	183	206	389 (58.2%)
Utterances that do not satisfy Con 1 but are segment boundary:	48	55	103 (15.4%)
Utterances with zero CBs :	30	121	151 (22.6%)
Utterances with more than one CB :	9	16	25 (3.7%)

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

	MUSEUM	PHARMA	TOTAL
Total number of realizations of CBs:	176	160	336
Total number of CBs realized as R1-pronouns:	88	43	131
CBs realized as personal pronouns:	88	43	131
CBs realized as relative pronouns:	0	0	0
CBs realized as demonstrative pronouns:	4	1	5
CBs NOT realized as R1-pronouns:	88	117	205
Total number of R1-pronouns that do not realize CBs:	52	19	71
Personal pronouns that do not realize CBs:	52	19	71
Relative pronouns that do not realize CBs:	0	0	0
Demonstrative pronouns that do not realize CBs:	3	12	15

The number of violations to Rule 1 also doubles again with respect to the u=s version with direct realization, from 12 to 25 (6.4% of the 389 utterances with a CB and a R1-pronoun). While this number of violations isn't enough to cast doubt on the validity of Rule 1, it is 3 1/2 times the number of violations with the 'Vanilla' version. The complete figures about violations and verifications for the three versions of Rule 1 are as follows.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	166	198	364 (93.6%)
GJW 95 - utterances that violate:	17	8	25 (6.4%)
Gordon - utterances that satisfy:	61	37	98 (25.2%)
Gordon - utterances that violate:	122	169	291 (74.8%)
GJW 83 - utterances that satisfy:	132	132	264 (67.9%)
GJW 83 - utterances that violate:	51	74	125 (32.1%)

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

The results concerning Rule 2 with the IS version are comparable to those obtained with the IF version; in particular, we get a large number of retaining transitions (115) and RSH (97) (as well as

104 EST, 64 CON, 34 SSH, 98 ZERO, and 156 NULL). Cheap transitions are 135 / 668, 20% of the total (as opposed to 14.7% with direct realization and 21.1% with IF), whereas 46.1% of transitions are expensive. The Kibble score is 1.95, much better than with u=s and direct realization (1.4) and IF (1.6).

Treating bridging references as containing null traces It might be thought that some of these additional violations of Rule 1 in versions IF and IS (such as the one in example (25)) shouldn't really be counted 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*:[37] these implicit anaphors might satisfy Rule 1. (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 R1-pronouns. I.e., the issue is the same raised by relative traces.) However, treating these null anaphors as R1-pronouns actually results in more violations of the rule, even though Rule 1 is still verified: from 14 to 23 with IF, and from 25 to 30 with IS. This is because although most bridging references do refer to the CB (see also (Sidner 1979; Poesio et al. 1998)), not all do, and every bridging reference not referring to the CB becomes a potential violation. In (26), 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.

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

On the positive side, this is the first parameter configuration among those discussed that verifies the version of Rule 1 proposed by (Gordon et al. 1993), both in the IF version (+345, -140, $p \leq .01$ by the sign test) and in the IS version (+253, -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 (from 21 to 94 (9.7%) with IF, from 25 to 87 (13%) with IS), as well as in the number of Rough Shifts (from 100 to 181 (18.6%) with IF, from 97 to 154 (23%) with IS). All in all, these results do not encourage us to adopt this proposal.[38]

Second Person CFs It has been suggested that second person pronouns (henceforth: PRO2s) introduce CFs, especially in dialogue (Byron and Stent 1998).[39] In the pharmaceutical domain, in particular, PRO2s are very numerous, and often seem to play an important role in maintaining the coherence of the discourse. In our corpus, allowing PRO2s to introduce CFs reduces 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 configuration), if we allow second person entities to count as CFs the statistics for the museum domain are not affected, but in the pharmaceutical domain the number of utterances that satisfy Strong C1 increases from 164 to 273, so that in total 466 utterances satisfy C1 and 364 violate it, which means that the constraint is verified by the sign test ($p \leq .03$).(The improvement is also significant: with 96 former violations being eliminated and only 5 new ones, $p \leq 0.01$.) With DS, 331 utterances verify the strong version of Constraint 1, and

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

[38] As we saw above, the same question comes up with traces in relative clauses, especially reduced relatives, which may also be argued to contain 'implicit' elements. In that case, as well, we find that it's best not to treat implicit anaphors as R1-pronouns.

[39] 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.).

214 violate it (as opposed to +264, -259 when second person entities are not treated as CFs). Allowing for indirect realization we get even better results for Strong C1: with IF, we get +623 and -241, a significant improvement even over the version with direct realization and PRO2s; with IS, +437, -145.

The results concerning Rule 2 are also improved by treating PRO2s as CFs. The percentage of NULL transitions is greatly reduced (for DF, down to 35% (from 47.7%); for DS, to 30% (from 40.9%); for IF, to 18.3% (29.7%); for IS, to 15.2% (from 23.3%)). As a result, the percentage of continuous transitions in Kibble's sense (EST, CONT, RET, SSH, RSH) increases, although RSH and SSH increase as well as EST and CONT. (In fact, in the IS version, RSH is now, with RET, the most frequent transition, at 18.3% each.) The overall figures for transitions in the IF version are shown in the following table.

	MUSEUM	PHARMA	TOTAL
Establishments:	74	121	195
Continuations :	50	77	127
Retain :	78	62	140
Smooth Shift :	35	36	71
Rough Shift :	61	58	119
Zero :	59	82	141
Null :	47	131	178

whereas for the IS version are:

	MUSEUM	PHARMA	TOTAL
Establishments:	45	52	97
Continuations :	28	63	91
Retain :	53	69	122
Smooth Shift :	9	34	43
Rough Shift :	57	65	122
Zero :	41	50	91
Null :	37	65	102

Finally, the Kibble coefficient increases for all versions: 1.51 for DF (vs. 1.14), 1.81 for DS (vs. 1.4), 1.95 for IF (vs. 1.6), and 2.3 for IS (vs. 1.95).

The results concerning Rule 1 crucially depend on whether we consider second person pronouns as R1-pronouns or not. Whether or not we do, letting second person entities introduce CFs results in more violations of Rule 1 (we concentrate here on the version from (Grosz et al. 1995)), both in absolute and in relative terms, because more utterances have a CB and therefore count as violations or verifications of the rule. But if we don't consider PRO2s as R1-pronouns, then the increase in violations is small: for DF, from 7 (2%) to 11 (2.7%); for DS, from 12 (4.5%) to 17 (5.1%); for IF, from 14 (3%) to 18 (3.2%); and for IS, from 25 (6.4%) to 30 (6.9%). If we do treat PRO2s as R1-pronouns, however, we find that the percentage of violations of Rule 1 almost triples for the u=f versions and doubles for the u=s ones: we now have 30 violations for DF (7.3%), 38 for DS (11.5%), 49 for IF (8.6%), and 66 for IS (15.1%). (Of course, Rule 1 still remains verified in a statistical sense in all of these cases.) The reason for this is that PRO2s do not seem to be very good indicators of the CB: about as many, or fewer, PRO2s occur as CBs as do not (for DF, 154 PRO2s refer to the CB, whereas 146 do not; for IS, 126 PRO2s refer to the CB, whereas 141 do not).

In the rest of the paper we will assume that second person entities introduce CFs, but are not R1-pronouns.

Predicative NPs The two changes to the definition of realization seen so far both had to do with increasing the number of CFs. What if we were to attempt to reduce the number of NPs 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 did not have any positive result at all; on the contrary, it did have a significant negative impact on Strong C1[40] 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*.[41]

(27) 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 treat predicative NPs as introducing CFs.[42]

Segmentation

As mentioned above, in the experiments discussed in this paper we didn't really study the effect of alternative claims about segmentation.[43] What we did compare were alternative heuristics for segmenting the text. Specifically, we looked at the differences that would result from having no segmentation at all, using major sections of a text as rough segments, and treating every paragraph as a separate segment. (See below, however.)

The basic (and obvious) result is that the smaller the segment, the better the results for C1, since utterances at segment boundaries are not counted as violations. The number of violations of Strong C1 increases progressively as segment size increases, and the constraint remains valid until the version in which sections are treated as segments. When an entire text is treated as single segments, Strong C1 only holds for IF and IS. Rule 2 in Grosz *et al.* version is unaffected 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, of course, 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, rather than only if it does not contain a pronoun referring to an entity in the previous paragraph (Walker's proposed heuristic) turned

[40]The difference is significantly worse for all the versions not treating PRO2s as CFs; worse, but not significantly so, if PRO2s are treated as CFs.

[41]This problem is tied with the issue of whether copular clauses should be uniformly viewed as asserting a predication, or if in some cases they can be viewed as stating an equality.

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

[43]Such as, say, (Walker 1998)'s proposal to replace Grosz and Sidner's focus space stack segments with a 'cache', or (Knott et al. ress)'s claim that (in certain genres, at least) segmentation is not entirely dependent on intentional structure.

out not to make any difference, since 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 (for Strong C1, we have +466, -405 (not significant); for the weak version, +856, -15, decrease: +23, -64.), DS, and IF. There was no difference with the IS configuration (+20, -20, the difference from the version using Walker's heuristic is not significant).

The results with no segmentation at all were significantly worse for all versions; the increases in violations go from 89 additional violations for the DF version, to -39 for the IS version. As a consequence, Strong C1 is not verified for the DF version, and it's only supported at the .04 level for DS.

Ranking Variants

Grammatical Function + Linear Disambiguation Because grammatical function does not uniquely specify a most highly ranked CF, some utterances end up having more than one CB, which causes the violations of the weak version of Constraint 1 seen above. However, this problem can be easily fixed by adding a tie-breaking factor. The most obvious choice for this, given, e.g., the results of Gernsbacher and Hargreaves (1988); Gordon et al. (1993), is linear order: so we might choose, e.g., the leftmost CF between two equally ranked CFs as having the highest rank. (We saw in Section §2 that linear order was already used by (Strube and Hahn 1999) to resolve tie-breaks, although they used a different ranking function.) It turns out that the results can also be slightly improved by ranking post-copular NPs in *there*-sentences (e.g., *someone* in *There is someone at the door*) as subjects rather than objects.

The resulting ranking function–henceforth abbreviated to GFTHERELIN–makes better predictions concerning local coherence as specified by Strong C1, irrespective of whether we identify utterances with finite clauses or sentences, and both with direct and indirect realization. With the DF configuration we have 481 utterances verifying Strong C1, and 349 violations (significantly better); with IF, +652, -212. The improvements are most significant for the u=s versions, since in sentences it's fairly common for more than one CF to be realized in the same grammatical position. The results for Strong C1 with GFTHERELIN and the DS configuration are +351, -194;[44] with IS, +475, -107 (37 utterances had more than one CB using normal grammatical function). The complete Strong C1 figures for IS are shown in the following Table.

	MUSEUM	PHARMA	TOTAL
Number of times at least one CF(Un) is realized in Un+1:	192	283	475 (71.1%)
Utterances that satisfy Constraint 1 (have exactly one CB) :	192	283	475 (71.1%)
Utterances that do not satisfy Con 1 but are segment boundary:	48	38	86 (12.9%)
Utterances with zero CBs :	30	77	107 (16%)
Utterances with more than one CB :	0	0	0

As in all previous cases, better results with Strong C1 are counterbalanced by worse results for Rule 1–although, again, not so much worse to result in R1 not being verified. The results with the DF configuration aren't significantly worse: +412, -12 for the strong version (as opposed to +398, -11). The complete results for all versions of Rule 1 with the DF configuration and GFHTERELIN are listed in the following table.

[44] These are the figures for the version including layout elements when necessary.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	150	262	412 (97.2%)
GJW 95 - utterances that violate:	7	5	12 (2.8%)
Gordon - utterances that satisfy:	87	50	137 (32.3%)
Gordon - utterances that violate:	70	217	287 (67.7%)
GJW 83 - utterances that satisfy:	134	182	316 (74.5%)
GJW 83 - utterances that violate:	23	85	108 (25.5%)

The results for R1 are significantly worse with the DS configuration: +329 (93.7%), -22 (6.3%) (versus +314 (94.9%), -17 (5.1%) with 'normal' grammatical function ranking). In two of the additional five violations of Rule 1, however, the problem is simply that by adding a disambiguation element we turn utterances whose CB is undefined (because more than one CF is equally ranked) into utterances with a CB. One such example is (28).

(28) (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 the IF configuration, the results are non-significantly worse, and are matched by an increase in the utterances that satisfy R1 (+577 (96.6%), -20 (3.4%) vs. +550 (96.8%), -18 (3.2%)). The full results for the three versions of Rule 1 under the IF configuration are as follows:

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	245	332	577 (96.6%)
GJW 95 - utterances that violate:	14	6	20 (3.3%)
Gordon - utterances that satisfy:	90	53	143 (24%)
Gordon - utterances that violate:	169	285	454 (76%)
GJW 83 - utterances that satisfy:	175	213	388 (65%)
GJW 83 - utterances that violate:	84	125	209 (35%)

Finally, the results with the IS configuration are also significantly worse at the .01 level (+439 (92.4%), -36 (7.6%) versus +407 (93.1%), -30 (6.9%) for the version with normal grammatical function ranking - a negative difference of 6). The overall figures for all three versions of Rule 1 under the IS configuration are shown in the following table.

	MUSEUM	PHARMA	TOTAL
GJW 95 - utterances that satisfy:	173	263	436 (92.4%)
GJW 95 - utterances that violate:	19	17	36 (7.6%)
Gordon - utterances that satisfy:	63	36	99 (21%)
Gordon - utterances that violate:	129	244	373 (79%)
GJW 83 - utterances that satisfy:	138	162	300 (63.5%)
GJW 83 - utterances that violate:	54	118	172 (36.5%)

In the case of Rule 2, the main change with GFTHERELIN is a strong reduction in the number of Rough Shifts, with all configurations (from 40-4.1%-to 29-3%-for DF; from 56 - 8.3%-to 44-6.6%-with DS; from 119-12.2%-to 98-10%-with IF; and from 122-18.3%-to 101-15%-with IS). With DF, we also observe minor increases in CON and a reduction in RET. The complete figures with this configuration are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	96	132	228
Continuations :	43	74	117
Retain :	25	30	55
Smooth Shift :	19	33	52
Rough Shift :	15	14	29
Zero :	66	84	150
Null :	140	200	340

The results for the DS configuration are similar: again, we find a small increase in CON and RET, and an even bigger decrease in RSH (from 56 to 44). With the IF configuration, again we have a small increase in CON and a decrease in RSH, but also a small decrease in RET. The overall figures for IF are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	74	121	195 (20%)
Continuations :	52	81	133 (13.7%)
Retain :	83	67	150 (15.4%)
Smooth Shift :	36	40	76 (7.8%)
Rough Shift :	53	45	98 (10%)
Zero :	59	82	141 (14.5%)
Null :	47	131	178 (18.3%)

Finally, for the IS configuration, we get again almost the same results concerning transitions, but with an increase in Smooth Shifts. The complete statistics about transitions for IS are as follows:

	MUSEUM	PHARMA	TOTAL
Establishments:	45	52	97 (10%)
Continuations :	30	67	97 (10%)
Retain :	56	73	129 (13.3%)
Smooth Shift :	10	41	51 (5.2%)
Rough Shift :	51	50	101 (10.4%)
Zero :	41	50	91 (9.4%)
Null :	37	65	102 (10.5%)

The change to GFTHERELIN hardly affects the relative percentages of cheap and expensive transitions; as for the Kibble score, it is increased under all configurations, but by a very small amount (e.g., from 1.95 to 1.99 for IF with PRO2s as CFs, and from 2.3 to 2.38 for IS).

IF and IS with GFTHERELIN ranking are clearly the best configurations using grammatical function as the basis for ranking; we will compare the configurations based on other approaches to ranking below to these two.

Linear Order Among the forms of ranking 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 were repeatedly observed in the literature by, among others, (Gernsbacher and Hargreaves 1988; Gordon et al. 1993; Stevenson et al. 1994).

It turns out that using this ranking function instead of GFTHERELIN doesn't result in significant differences. This is easy to understand in the case of Constraint 1, since all that matters for the

constraint to be verified is whether discourse entities are mentioned in successive utterances, and whether the ranking function always results in a single most highly ranked entity. However, we didn't observe any significant differences as far as Rule 1 is concerned, either, although the version using linear order as the ranking function consistently performs slightly worse than its corresponding version with GFTHERELIN. With DF, we find two additional violations when using linear order as a ranking function, but one of the previous violations disappears, which we summarize as +1, -2. With DS, the results of the comparison are +2, -3; with IF, +1, -3; and with IS, the results with linear order are exactly equivalent to those with GFTHERELIN and we have a tie - +4, -4.

Linear order also results in slightly worse results as far as Rule 2 is concerned, in that a few moves previously classified as continuations become retains (2 with DF and DS, 6 with IF and IS) and a few Smooth Shifts become Rough Shifts (again 2 with DF and DS, 1 with IF and IS). The Kibble score also gets very slightly lower throughout (e.g., for IS, from 2.38 with GFTHERELIN to 2.36 with linear order). All in all, these results do not suggest that linear order is a better ranking than GFTHERELIN; however, it might be advantageous in some cases, since it is easier for applications to compute.[45]

Combining Grammatical Function and Linear Order The experiments by Gordon et al. (1993) suggest that subjecthood and first-mentionhood result in equal ranking for CFs. 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. 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 configuration and ranking=GFTHERELIN, but not with the combined ranking, is the following, in which *Sieber-Fuchs* is pronominalized in (u2).

(29) (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.

(Notice that *Sieber-Fuchs* is the CB in this case because of the added order-based disambiguation.) The results for R1 with the four configurations are as follows: with DF, +413, -11 (+1); DS, +329, -22, +1, -1; IF, +578, -19, +1; IS, +436, -36 (identical). The results concerning Rule 2 with this configuration 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 CONT become RET, and a few others change from SSH to RSH. With the DF configuration we only have a change from CON to RET: CON=112 out of 971, 11.5% (instead of 117 with GFTHERELIN) and RET=61 (instead of 55). With DS we see the same change, but also one from SSH (34 vs. 40) to RSH (47 vs. 44). IF is like DF: CON=124 (133 with GFTHERELIN), RET=158 (was 150), SSH and RSH remain the same. With IS, CON goes from 97 / 668 (14.5%) to 91 (13.6%), RET=134 (was 129), SSH=46(was 51), RSH=107 (was 101). The Kibble scores are all very slightly lower: KS=1.51 for DF (down from

[45] On the other hand, a series of results by Gordon and collaborators (e.g., (Gordon et al. 1993, 1999)) suggest that position in the syntactic tree is a better predictor of salience than linear order. Prasad and Strube (2000) also found that in Hindi, linear order doesn't seem to have an effect on ranking.

1.54), KS=1.84 for DS (1.86), KS=1.95 for IF, KS=2.35 for IS. The percentage of cheap transitions is also lower throughout–e.g., with IS we have Cheap=171, Expensive=272 (vs. 175 and 268).

Informational Structure We didn't expect the results for Strong C1 to change by replacing GFTHERELIN with the ranking function based on information structure proposed by Strube and Hahn (1999), for the same reasons as we didn't expect them with linear-order, and indeed we didn't find any. Less expected was the fact that we didn't find any significant differences as far as Rule 1 is concerned, either (again, just as in the case of linear order). (We only discuss here the results with the version from Grosz et al. (1995).) With the DF configuration we have 414 utterances verifying R1 and 13 violating it: 2 violations with GFTHERELIN now verify the rule, and 1 new violation. (This difference is not significant.) With DS, we have +332, -19 vs. +329, -22 with GFTHERELIN (+5, -2, again not significant.) On the other hand, with IF we get worse results for Rule 1 than with GFTHERELIN, although again these differences are not significant: +577, -23 vs. +577, -20 (+1, -4). Finally, with IS we have quite a lot of differences, but the overall results are identical to those obtained with GFTHERELIN: 436 utterances verify the rule, 36 violate it.

The one claim of Centering Theory where we can find a difference between the ranking function proposed by Strube and Hahn and GFTHERELIN is Rule 2. Although we do not know of statistical tests that can back up this impression, with all four configurations we have been considering, replacing GFTHERELIN with the Strube-Hahn ranking function results in more continuations and fewer retains, more smooth shifts and fewer rough shifts (although NULL, ZERO and EST remain the most common transitions); more cheap transitions, and fewer expensive ones (although we still have more expensive than cheap transitions with all configurations); and higher Kibble scores. With DF, we still have NULL, EST, and ZERO as the three most frequent transitions, and in about the same proportions as with GFTHERELIN (35%, 23.5%, 15.4%); but we also have more continuations (CON=141 / 971, 14.5% (vs 117 with GFTHERELIN)) and fewer Retains (RET=33, 3.4% (55)); more Smooth Shifts (SSH=56, 5.7% (52)), and fewer Rough Shifts, RSH=23, 2.4% (29). We still find more expensive transitions than cheap ones (EXP=518, CHP=228), but the percentage of cheap transitions is slightly better (23.5% vs. 21.3%); and the Kibble score is higher, KS=1.65 (vs. 1.54).

The same happens with the other three configurations. With DS, we have a similar reverse between continuations and retains (CON=123, 18.4% (94) and RET=32 (58)), but with Strube / Hahn ranking, unlike with GFTHERELIN, we also have more SSH (56, 8.3% (40) than RSH (25, 3.7% (44)). We have the same slight improvement in the number of cheap transitions as with DF (CHP=153, EXP=290, vs. 137 and 306) and a higher Kibble score, KS=2.08 (vs. 1.86). With IF, we still have more SSH than RSH (SSH=82(76), RSH=79(98)) and, in addition, we also have more CON than RET: CON=160 (133), RET=136(150). And again, we have a small improvement in the relative numbers of cheap transitions and in the Kibble score (CHP=288, EXP=458; KS=2.13 (vs. 1.99)). And finally, with IS– which is the closest configuration to the one proposed by Strube and Hahn, apart from our inclusion of second person entities, we have CON=125, 12.8% (97) and RET=112 (129); about the same SSH and RSH, SSH=69 (51), RSH=72(101); the higher percentage of cheap transitions (CHP=203, 30%, and EXP=240); and the highest Kibble score obtained by any configuration, 2.62 (vs. 2.38). Note however that even in the 'best' version, it's still the case that we have more expensive transitions than cheap ones, and more EXP-EXP than CHP-CHP sequences, contrary to Strube and Hahn's version of Rule 2: 79 vs. 108 with IS, the configuration closest to the one studied by Strube and Hahn.[46]

[46] In (Poesio et al. 2000) only Constraint 1 and Rule 1 were studied, and therefore the ranking function proposed by Strube and Hahn was found to be equivalent to GFTHERELIN. The new results suggest that the S&H ranking function leads to results more agreement with the various versions of Rule 2 than GFTHERELIN, although it's not

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 configuration, using Gordon *et al.*'s definition of CB and the ranking function they propose in that paper, there are 147 more violations of Strong C1; however, the number of violations of R1 goes down from 12 to 5, also a significant improvement (-8, +1, $p \leq 0.02$). Most of these are simply 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 (29), already seen above, where *she* in (u2) is now the CB. Another case is (30). Because the CB has to be the subject, the fact that 'Louis XIV' had higher ranking in the previous utterance doesn't matter; the only possible CB is *he*.[47]

(29) (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 .

(30) (u306) In 1672, the king granted him the royal privilege of lodging in the Palais du Louvre.

(u307) In the same year, he achieved the title of cabinetmaker and sculptor to Louis XIV, King of France.

On the other hand, the utterance which immediately follows (u307) in the same text as (30), (u311) (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.

(31) (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, it appears that 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 IF configuration illustrates another characteristic of this configuration: using Gordon *et al.*'s definition of CB results in a virtual elimination of all types of transitions apart from continuations and establishments. We find 5 RET, 20 SSH, and only 1 RSH–this is the version that results in the fewest 'incoherent' transitions. The reduction in the number of violations of Rule 1 is even greater for the u=s configurations. With DS we have only 8 violations, 11 fewer than the version using the 'vanilla' definition of CB. Even larger reductions in the number of violations of R1 are found with the IS configuration, down to 8, from 36 with the 'classic' definition in Constraint 3 and GFTHERELIN.

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 configuration can be summarized as follows: very few utterances

clear to us whether these results are significant.

[47] As discussed above, one might argue that in the second utterance 'Louis XIV, king of France' is not really used to refer to the individual, but is part of the title. It's not clear however whether such distinctions should be made in Centering Theory and / or whether our annotators would be able to make them.

end up having a CB (more precisely, a Local Center); but once it is established, nothing else gets pronominalized.

With the DF configuration and GFTHERELIN for ranking, for example, only 20 utterances have a CB, but we have no Rule 1 violations at all, for any of the versions of R1 that we considered. An example in which the CB / Local Center does get established (in (u64)) is the following:

(32) (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.

On the other hand, the link between pronominalization and 'center' seems to be completely lost in this version. While it is true that 23 out of 24 realizations of a CB in this case are done via pronouns, it is also true that 194 personal pronouns are not realizations of CB; so a separate story will be needed to account for the cases (the great majority) of pronouns not referring to the Local Center.

In terms of transitions, we have 19 establishments, 1 continuation, 16 zero, 935 nulls, and 0 everything else: i.e., no shifts, and no retains.

The same pattern is encountered with the other utterance / realization configurations. With DS, only 18 utterances have a CB, and all satisfy R1 (this is 22 fewer violations of R1). We only have 34 realizations of a CB, of which 27 done via pronouns; 190 pronouns are not realizations of CBs. With IS, 18 utterances have a CB, 427 don't have one, so the comparison on C1 with the version using C3 as definition of CB is +0, -316. On the other hand, on R1 we have 36 fewer violations, and no new ones.

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 OTHER CLAIMS ABOUT DISCOURSE CAST IN TERMS OF CENTERING

The claims of Centering Theory analyzed so far, and especially Constraint 1, and Rule 2, are primarily claims about the 'building blocks' of the theory. Already Rule 1 is more of a 'linguistic' claim, in that notions of the theory are used to predict a linguistic phenomenon (the form of an NP); and we saw in Section §2 that the concepts of Centering Theory have been used to make other claims of this kind–e.g., abot the correlation between centering transitions and the form of the subject, or the type of discourse entities that may serve as the antecedents for long-distance pronouns. These data can also be very useful to identify the 'best' parameter configuration: presumably, the 'best' configuration will be the one which makes more useful predictions. In this section we return to these claims in the light of the results just presented concerning the 'best' ways of setting the parameters of Centering.

Type of Transition vs Form of Subject

Kameyama (1986); Di Eugenio (1998); Turan (1995) 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' configurations identified above. However, because of the low frequency of some events,[48] our results should be considered as preliminary.

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

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	7	1	79
CON	21	2	62
RET	2	3	94
SSH	11	0	33
RSH	3	3	84
ZERO	1	2	50
NULL	1	1	49
TOTALS:	46	12	451

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. An obvious 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, with 5 degrees of freedom, and with no cells with zero elements:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON	21	64	85
RET	2	97	99
RSH-SSH	14	120	134
ZERO	1	52	53
NULL	1	50	51
TOTAL	46	463	509

For this table, $\chi^2 = 38.1, p \leq 0.001$, a possibly significant result; but this table still contains cells with values under 5, which tend to increase the χ^2 value, so more drastic collapses 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). The resulting contingency table is as follows:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON-RET	23	161	184
RSH-SSH-ZERO	15	172	187
NULL	1	50	51
TOTAL	46	463	509

[48] 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.

This distribution however is not significant: with 3df, we have $\chi^2 = 6.13$, whereas for significance at the .05 level, we need $\chi^2 \geq 7.82$. Perhaps mixing CON and RET is not a good idea, as indeed one might suspect from the fact that whereas 1 in 4 CON is signalled by a pronoun, the percentage for RET is much lower (almost 2 in 100). Another way of eliminating the low counts is to simply drop RET and NULL, while maintaining the merge of ZERO and SSH-RSH:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON	21	64	85
RSH-SSH-ZERO	15	172	187
TOTAL	43	316	359

This new table also doesn't have low count cells, and the distribution this time is highly significant: $\chi^2 = 17.1, p \leq 0.001$. An alternative way to get rid of the low counts is to just ignore ZERO and NULLs, keeping RET distinct from CON:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON	21	64	85
RET	2	97	99
RSH-SSH	14	120	134
TOTAL	44	361	405

With this contingency table, as well, the dependence of the two variables is quite strong: with 3df, $\chi^2 = 25.6, p \leq 3x10^{-5}$. However, because this table contains one low count cell, the solution above looks more preferrable.

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, rather than CON and RET. On the other hand, 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. This skepticism seems to be confirmed by our results. The contingency table is as follows:

	PRONOUN	FULL NP	TOTAL
EST / CON	28	144	172
RET	2	97	99
SH	14	120	134
TOTAL	44	361	405

With this table, if we collapse EST and CON we still get a dependency between the two variables, but lower: with 2df, $\chi^2 = 13.2, p \leq 0.002$. And significance completely disappears if we eliminate the low-count RET line:

	PRONOUN	FULL NP	TOTAL
EST / CON	28	144	172
SH	14	120	134
TOTAL	42	264	306

Now $\chi^2 = 2.16, p \leq 1$, whereas with 1df, χ^2 should be greater than 3.84 for significance at the .05 level.

Finally, one might think of an even simpler two-way distinction, between continuations and shifts, treating EST as a type of SHIFT. The resulting distribution is shown in the following contingency table:

	PRONOUN	FULL NP	TOTAL
CON	21	64	85
RSH-SSH-ZERO-EST	22	252	274
TOTAL	43	316	359

This distribution is again highly significant (1df, $\chi^2 = 17.1, p \leq .001$), about as much as the one with a three-way distinction between EST, CON and RSH-SSH-ZERO. But again, the alternative merging of CON and RET, as in the contingency table below, is not significant:

	PRONOUN	FULL NP	TOTAL
CON-RET	23	161	184
RSH-SSH-ZERO-EST	22	252	274
TOTAL	45	413	458

Finally, merging CON with EST results in a distribution that is still significant, but only at the .05 level:

	PRONOUN	FULL NP	TOTAL
CON-EST	28	144	172
RSH-SSH-ZERO	15	172	187
TOTAL	43	316	359

IF, GFTHERELIN**:** If we consider instead the IF configuration and GFTHERELIN ranking, we get the following contingency table:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	21	1	131
CON	40	2	76
RET	0	0	109
SSH	18	0	23
RSH	5	6	75
ZERO	1	3	89
NULL	5	3	82
TOTALS:	90	15	585

Collapsing RSH and SSH, and the two columns DEM and FULL, is again not enough to completely eliminate the low counts:

	PRONOUN	FULL NP	TOTAL
EST	21	132	153
CON	40	78	118
RET	0	109	109
RSH-SSH	23	104	127
ZERO	1	92	93
NULL	5	85	90
TOTAL	90	600	690

The high $\chi^2 = 80.7$ (with 5df, $p \leq 0.001$) for this distribution is therefore rather dubious. Eliminating RET, and merging ZERO with the shifts (no need to eliminate NULLs in this case), we get a contingency table with sufficient counts in all cells:

	PRONOUN	FULL NP	TOTAL
EST	21	132	153
CON	40	78	118
RSH-SSH-ZERO	24	196	220
NULL	5	85	90
TOTAL	90	491	581

For this table, $\chi^2 = 41.2$, which with 3df is highly significant. Eliminating NULLs we get a distribution with the same .001 degree of significance as the equivalent one with IS ($\chi^2 = 30.4$):

	PRONOUN	FULL NP	TOTAL
EST	21	132	153
CON	40	78	118
RSH-SSH-ZERO	24	196	220
TOTAL	85	406	491

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	7	1	79
CON	22	5	87
RET	1	1	78
SSH	13	1	48
RSH	1	1	60
ZERO	1	2	50
NULL	1	1	49
TOTALS:	46	12	451

Collapsing RSH and SSH, and the columns DEM and FULL, but keeping ZEROs and NULLs, again it's not sufficient to completely eliminate low count cells:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON	22	92	114
RET	1	79	80
RSH-SSH	14	110	124
ZERO	1	52	53
NULL	1	50	51
TOTAL	46	463	509

So again we drop RET and NULL, and merge RSH, SSH, and ZERO:

	PRONOUN	FULL NP	TOTAL
EST	7	80	87
CON	22	92	114
RSH-SSH-ZERO	15	162	177
TOTAL	44	334	378

This distribution is significant at the .01 level ($\chi^2 = 9.32$).

STRUBE-HAHN, IF: The full contingency table for this configuration is as follows:

	PERS PRONOUN	DEM PRONOUN	FULL NP
EST	21	1	131
CON	40	2	99
RET	1	1	94
SSH	20	1	37
RSH	2	4	53
ZERO	1	3	89
NULL	5	3	82
TOTALS:	90	15	585

Collapsing as above, we get the following distribution, significant at the .001 level ($\chi^2 = 19.8$).

	PRONOUN	FULL NP	TOTAL
EST	21	132	153
CON	40	101	141
RSH-SSH-ZERO	23	187	210
TOTAL	84	420	504

Summary We observed 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). The dependency is significant for all four configurations shown to be 'best' by the analyses in Section §4. 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.

Our results also suggest that at least 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); from this point of view, establishments seem to pattern more with shifts. Establishments are best grouped with shifts than with continuations also when a two-way classification is considered.

The correlation between transitions and segment boundaries

Another use of notions from Centering theory to analyze (discourse) linguistic behavior was considered in (Walker 1998; Passonneau 1998), who 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 considered one configuration of the theory; so we tried to see if we could get a better result using the 'best' configurations 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 configuration is shown in the contingency table below:

	NOT BOUNDARY	BOUNDARY	TOTALS:
EST	140	55	195
CON	115	18	133
RET	129	21	150
SSH	65	11	76
RSH	85	13	98
ZERO	91	50	141
NULL	121	57	178
TOTALS:	746	225	971

It is obvious from the table that the correlation between CB continuation and segment continuations is imperfect at best; there is, however, a certain tendency, confirmed by the results of the χ^2 test, which give very high results–with 6df, $\chi^2 = 45.22, p \leq .001$.

An obvious observation about the table above is that CON and RET are less frequently boundaries than SSH, RSH and ZERO. Notice also that in this case, just as in the case of the correlation between transition and subject type, establishments are rather different from continuations: about 1/4 of EST are boundaries, whereas only 1/10 of CON are. The other transition that correlates relatively more highly with boundaries is NULL (1/3 of NULL are boundaries). In fact, EST and NULL are more frequently boundaries that the shifts or ZERO. This suggests collapsing the categories as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	244	39	282
SSH+RSH+ZERO+EST+NULL	502	186	689
TOTALS:	746	225	971

This table makes the correlation very obvious: boundaries are 14% of the total number of CON+RET transitions, but as frequent with the other class of transitions. This is confirmed by the results for the χ^2 test (with 1df, $\chi^2 = 19.78, p \leq 0.001$).

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

	NOT BOUNDARY	BOUNDARY	TOTALS
EST	56	41	97
CON	73	24	97
RET	98	31	129
SSH	34	17	51
RSH	75	26	101
ZERO	49	42	91
NULL	58	44	102
TOTALS:	443	225	668

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, $\chi^2 = 25.32, p = 3x10^{-4}$). A collapsed table again makes the correlations more obvious (boundaries are 24% of CON+RET, but 38% of the rest) although the χ^2 value, 13.3, is lower than with IF.

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	171	55	226
EST+SSH+RSH+ZERO+NULL	272	170	447
TOTALS:	443	225	668

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

	NOT BOUNDARY	BOUNDARY	TOTALS
EST	140	55	195
CON	139	21	160
RET	115	21	136
SSH	71	11	82
RSH	69	10	79
ZERO	91	50	141
NULL	121	57	178
TOTALS:	746	225	971

This distribution is highly unlikely to be due to chance, just like the one with GFTHERELIN: $\chi^2 = 45.49, p \leq 0.001$. The 'collapsed' distribution is as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	254	42	296
EST+SSH+RSH+ZERO+NULL	492	183	675
TOTALS:	746	225	971

Again, given this contingency table it is highly unlikely that the two variables are independent, $\chi^2 = 19.3, p \leq 0.001$. Both of these χ^2 values are virtually identical to those obtained with GFTHERELIN.

STRUBE-HAHN, IS: 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	56	41	97
CON	94	31	125
RET	82	30	112
SSH	52	17	69
RSH	52	20	72
ZERO	49	42	91
NULL	58	44	102
TOTALS:	443	225	668

This table has $\chi^2 = 24.07, p \leq 0.001$. The collapsed table is as follows:

	NOT BOUNDARY	BOUNDARY	TOTALS
CON+RET	176	61	237
EST+SSH+RSH+ZERO+NULL	267	164	431
TOTALS:	443	225	668

The χ^2 value for this table, 10.37, is still significant but only at the 1% level.

Summary Our χ^2 tests 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' configurations and the collapsed classes discussed above (CON+RET, EST+ZERO +SSH+RSH+NULL) we get results comparable to Passonneau's. With IF settings and GFTHERELIN ranking (the configuration that results in the highest χ^2 value for the correlation between transitions and boundaries), and using the class EST+SSH+RSH +ZERO+NULL to predict boundaries, we get R=.83, P=.27, E=39+502/971 = .55. Using CON+RET to predict nonboundary, R=.33,P=.86, and E stays the same. We get slightly worse results using Strube-Hahn ranking and IS: using EST+SSH+RSH +ZERO+NULL to predict boundary, we have R=.73, P=.38; using CON+RET to predict non-boundary, R=.39, P=.74. In other words, even by using the 'best' configurations 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.

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 configurations.[49]

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 about 18 pronouns out of 217 are long distance, which is the same percentage (8%) found in the corpus used by Hitzeman and Poesio.[50] If we identify utterances with sentences, however, we only get 8 long-distance pronouns.[51]

Hitzeman and Poesio's claim is verified in our corpus as well, both for the IF configuration and for the IS configuration. With IF, 15 long distance pronouns out of 18 had been CBs and 3 had not, $p \leq .04$ both with GFTHERELIN ranking and with Strube / Hahn ranking. With IS, we find +6, -2 with GFTHERELIN ranking, and +7, -1 with Strube / Hahn ranking, but there is not enough data for a significance test. An even better result was found however with the IF configuration by weakening the licensing condition to having been a CP rather than a CB: in this case, with IF we have +17, -1, $p \leq .01$ by the sign test, with both GFTHERELIN and Strube-Hahn ranking. (With IS, the results are the same as for the case in which the discourse entity had occurred as a CP.)

Fronted Clauses: Subordination vs. Linearization

We conclude this section by examining a further variant in the definition of previous utterance. As discussed above, Suri and McCoy proposed that an adjunct clause *at the end* of a sentence is treated as embedded, i.e., it is 'passed over' when looking for the previous utterance of a following clause, much like Kameyama proposes it's the case for complement clauses. Suri and McCoy however do not discuss the case in which the subordinated adjunct is the *first* clause in the sentence, which happens, e.g., with conditional clauses (*If John wants to have dinner, he'd better get home quickly*). The results we have discussed so far were obtained by treating such clauses 'Kameyama-like': e.g., in the following example, (u1) would be treated as previous utterance for (u2), but then (u2) rather than (u1) would be the previous utterance for (u3).

(33) (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.

There are 51 such clauses in our corpus. We considered what would happen by treating such clauses as embedded, as well (i.e., by viewing (u1) as previous utterance for (u3), instead of (u2)). With GFTHERELIN ranking and IF setting, the results are clearly worse, especially for C1: we now find 249 violations to the strong version of the constraint (+615, -249), as opposed to just 212 for the 'linearized' version; the difference is significant. We do have 2 fewer violations of Rule 1, but this difference is not significant. There are no differences between GFTHERELIN and STRUBE-HAHN ranking; and of course there are no differences with the IS setting.

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

[50] 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.

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

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 (34), but not in (35), where the same underlying semantic subordination is not syntactically realized:

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

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

In subsequent work, 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.

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).

Figure 1 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.

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.

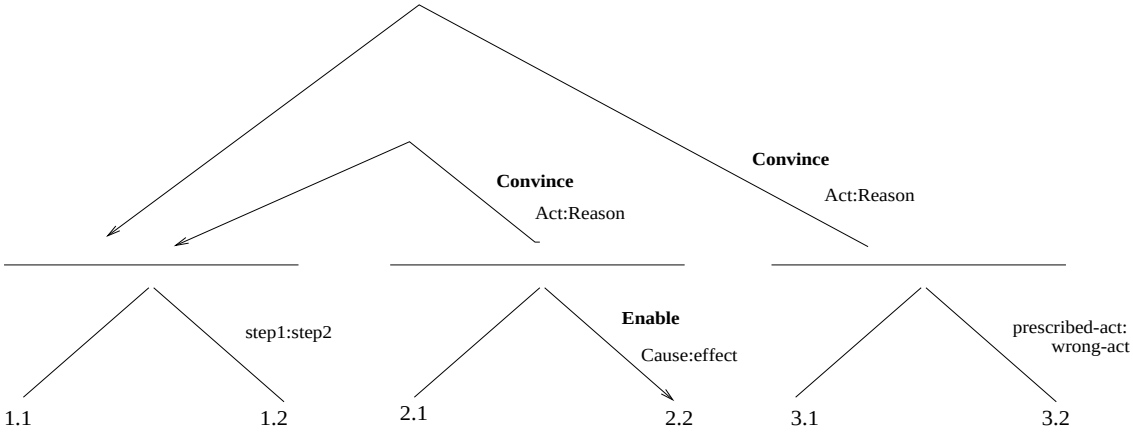

Figure 1: A tutorial excerpt and its RDA analysis

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.

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).

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.

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 1, 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 1 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.

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 configuration 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 one discussed previously and making claims about sequences of transitions, and one which simply says that continuations are preferred over retains, which in turn are preferred over shifts (see, e.g., the introduction to (Walker et al. 1998b). 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.

Discussion

This experiment confirms one of the findings of the other experiments: 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. 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.

(36) 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

A comparison between different versions of the theory

One of our goals was to compare the many different ways of instantiating Centering Theory on a single data set. The first result of this analysis is that in what we called the 'vanilla' version only Rule 1 is verified (except for the version of the Rule proposed by Gordon *et al.*). The strong version of Constraint 1 clearly isn't verified; as for Rule 2, the story is complicated, but one could certainly find the presence of so many NULL transitions surprising, as well as the fact that if we count SSH and RSH together, there are more shifts than retains, and as many as there are continuations. Another unexpected property of this version is the existence of utterances with more than one CB.

On the other hand, we saw that there are several ways of fixing the parameters of the theory so that Strong C1 is satisfied. The two choices of parameters with the most dramatic impact on Strong C1 are allowing for indirect realization and identifying utterances with sentences. Either one of these changes by itself is sufficient for Strong C1 to be verified. If in addition we use a ranking function with a disambiguation component, like GFTHERELIN, the multiple CB cases are eliminated, as well. The resulting configurations, which we have called IF and IS, verify both Strong C1 and the two 'basic' versions of R1. With these two configurations, however, we find more Rough Shifts than Smooth Shifts, and more Retains than 'pure' Continuations (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); as well as in many more expensive transitions than cheap ones.

Changing the ranking function from GFTHERELIN to Strube-Hahn does not affect the results for Strong C1 or R1 at all, but it does result in a distribution of transitions which is closer to what one would expect on the basis of most versions of Rule 2; in particular, we get more SSH than RSH. So, if we were looking for the best 'all-rounder' configuration, that would be either IF or IS with Strube-Hahn ranking (and treating second person NPs as introducing CFs). Unfortunately it's not clear whether the differences concerning Rule 2 are significant; and it would be nice to see the positive corpus-based results obtained with this type of ranking supported by psychological research in the way grammatical function ranking has been (Hudson et al. 1986; Gordon et al. 1993; Brennan 1995). Also, we should point out that even these configurations don't support Strube and Hahn's own version of Rule 2, but only those proposed by Brennan *et al.* and by Kibble.

We should also remember that a third major result of this work is that at least with our corpus, talking of a 'best' version is not completely correct, because improvements in one direction tend to lead to worse results in the other. In particular, we saw that there is a clear tradeoff between Strong C1 and R1: reducing the number of violations of Constraint 1, whether by identifying utterances with sentences, or by allowing indirect realization, results in significant increases in the number of violations to Rule 1, although at least two of the versions of the Rule are so robust that they still hold even in these versions. (Yet, in the IS version, 7% of the utterances that contain a pronoun violate the principle.) Perhaps the most spectacular demonstration of this tradeoff are the versions of the theory that adopt the definitions of CB proposed by Gordon et al. (1993) and Passonneau (1993). By adopting a particular restrictive definition of CB, these versions succeed in reducing (indeed, eliminating, in the case of Passonneau) the number of violations of R1 - but the price is that only very few utterances have a CB.

These changes also affect R2. Both the identification of utterances with sentences and indirect realization result in a great increase in the number of RSH, that become more common than SSH; we also find more RET than 'pure' CON.

Both 'utterance' and 'realization' are clearly key parameters of the theory, whose definition has to

be considered very carefully; in neither case the choice should be led only by the desire to minimize the violations to the claims. In part this is because of the tradeoff just discussed. In the case of the definition of utterance, there are some reasons for preferring clauses to sentences: e.g., most analysis of discourse (e.g., Rhetorical Structures Theory (Mann and Thompson 1988)) view clauses as the basic unit of discourse. Our results do however support Kameyama's idea that if clauses are viewed as the unit of local focus update, only finite ones should be considered; treating all clauses as utterances results in significantly worse results.

Other alternative definitions of the parameters of the theory do not make much difference, or have a negative impact. Especially surprising, given the importance given to the issue in the literature on Centering, is the fact that alternative ranking functions–linear order, or a combination of grammatical function and linear order–did not result in significant differences. Even the ranking function proposed by Strube and Hahn only affects the classification of transitions, an aspect of the theory which has yet to prove of empirical significance (see the discussion in Section §5). Changes in the definition of previous utterance also have limited impact. Adopting a 'Suri-like' notion of which utterance should be chosen as previous in cases of adjunct clauses results in fewer violations of C1 than with Kameyama's, but not so many that C1 is verified, and only if we treat relative clauses 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.

We should note however that some of the alternative definitions of the parameters are supported by empirical evidence such as psychological results, that should supplement our results. In the case of the definition of previous utterance, we saw that psychological experiments support a 'Suri-like' approach, 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 ranking, whereas grammatical function ranking and linear order leads to undistinguishable results for English, Prasad and Strube's work on Hindi (2000) indicates that in more free-order languages the difference may be significant; so do Strube and Hahn's results concerning grammatical function vs informational structure in German. (It would be interesting to compare STRUBE-HAHN with GFTHERELIN in Hindi.)

We also saw that some choices not seriously discussed in the literature turned out to have a significant impact. One such issue is the definition of 'R1-pronoun': i.e., whether we should consider traces in relative clauses, the implicit anaphoric elements of bridging references, or demonstrative pronouns, among the 'pronouns' to which Rule 1 applies, or not. Another important issue is the treatment of second person entities. Our results indicate that if we do not treat second person entities as introducing CF, or we treat PRO2s as R1-pronouns, there are many more violations to Strong C1 and R1, respectively (although the principles are still verified). Just as in the case of relative clauses, we have a dilemma: whether to just consistently choose the version that results in fewer violations of the claims, possibly at the cost of adopting dubious theoretical positions; or if instead sometimes we should accept more violations, and leave the task of accounting for such cases to other aspects of discourse. Two other such cases are the choice between finite clauses and sentences as utterances, and the treatment of noun phrases in non-referring positions.

Is Centering Theory verified or not?

A more fundamental question addressed in this work is whether at least one among the variants of the theory we discussed is such that the claims of Centering Theory are verified. In order to answer this question, however, we have to decide how these claims should be interpreted.

If we were to interpret them as strict rules that admit no exceptions, in the manner of linguistic claims such as the Conservativity principle for quantifiers (Keenan and Westerståhl 1997) or Principle

A of the binding theory, then we would have to conclude that none of the claims of the theory is verified, as we found exceptions to all of them. This is not, however, how claims such as Constraint 1 or Rule 1 were meant to be interpreted; rather, they are clearly stated as preferences which, when satisfied, make discourses easier to read. Therefore, we tested these claims in a statistical sense, by means of significance tests. We have seen than when the theory's claims are interpreted in this way, it is possible to set the parameters of the theory in such a way that at least C1 and R1 are clearly verified, and possibly R2 as well, although we also saw that R2 is weaker. We discuss each principle in turn.

CT as a theory of pronominalization (and salience) The more robust of the theory's claims are those about pronominalization and salience, expressed by Rule 1. We saw that at least two of the versions of R1 proposed in the literature are verified under pretty much all parameter configurations, and in a very convincing way: of the variants in which both C1 and R1 are verified, the utterances violating R1 are between 3% of the total for the direct realization versions and 8% for the ones with indirect realization.

On the other hand, R1 is a very weak claim, that couldn't really be used as the basis for a theory of pronominalization (Henschel et al. 2000). All it says is that *if* we decide to pronominalize, *then* we should pronominalize the CB –but this formulation doesn't address the real problem for NLG systems or for theories of production, which is to find when is it that one should pronominalize. The statistics about pronominalization presented in the paper, as well as the poor showing of Gordon *et al.*'s version of Rule 1, at least with the variants that also verify Constraint 1, indicate quite clearly that CBs are not pronominalized as frequently as one would imagine: less than half of CB realizations are via pronouns. The opposite is also true: e.g., in the IS version with GFTHERELIN ranking, 86 out of 217 third person personal pronouns refer to non-CBs. Examples like (19) illustrate one reason why this happens: a discourse entity may be sufficiently salient to justify pronominalization by having been referred in the text often. This discrepancy between the theory's predictions and our data is analyzed 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% on the museum domain.

CT as a theory of coherence: Constraint 1 One of the main results of this work is that the validity of Centering's claims concerning local coherence depends to a significant extent on the choice of the parameters, much more so than in the case of the claims about local salience and pronominalization. This applies for both Constraint 1 and Rule 2. Specifically, Strong C1 does not hold for what we called the 'vanilla' version of the theory. While it is true that this parameter configuration is a bit of a straw man in that it has never been explicitly proposed in this form, we do believe that the choices adopted in this version are those most researchers outside the area would associate with the theory. Strong C1 does however hold for any version which either identifies utterances with sentences or allows for indirect realization. (While the weakest version of C1–only requiring that there is at most one most salient entity per utterance– does hold even for the vanilla version, it's not quite as interesting as a claim about coherence. And anyway, we found quite a few counterexamples to this version, as well.)

Even in the best case, there are many more exceptions to Strong C1 than we found for Rule 1 (between 20 and 25% of the total number of utterances) even when adopting a pretty fine-grained notion of segment. Assuming that our texts were coherent, this suggests to us that there must be other ways of achieving local coherence, apart from what we have been calling here 'entity coherence'.

An obvious candidate are rhetorical relations; indeed, since the very beginning of discourse analysis (Kintsch and van Dijk 1978; Hobbs 1979) there has been a feeling that 'entity' coherence needs to be supplemented by 'relational' coherence. This hypothesis is supported by an analysis of our data.

One case of violations to Constraint 1 in the museum domain are utterances 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 (37), (u2) gives background concerning the decoration of a cabinet. In (38), utterances (u2)-(u5) give information about a particular class of rings to which the objects under discussion belong. Note that whereas in the case of (u1)-(u2) one may conceivably treat *poligonal openwork rings* as a bridging reference to *two gold finger-rings*, in the case of (u3) it is more difficult to find a clear bridging reference.[52]

(37) (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.

(38) (u1) <title> Two gold finger-rings from Roman Britain (2nd - 3rd century AD). </title> (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 panels are in Greek: That on 602 (left), from Corbridge, Northumberland, reads: 'the love-token of Polemios'.

While it is true that some of these violations could be fixed by adopting a broader notion of bridging reference–e.g., in (37) we might treat *France* as a bridging reference to *Louis XIV*–we are skeptical that this wider notion of bridge can be annotated reliably.

The pharmaceutical leaflets contain many examples in which the connection between clauses is explicitly indicated by connectives, as in (39), repeated here:

(39) (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 many such cases, letting second person pronouns introduce discourse entities results in them being classified as 'entity-coherent' even though one may think that the coherence is actually achieved by way of the explicit indication of the rhetorical relation. One example is (40).

(40) 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?

In fact, some might argue that we don't really need a notion of 'entity coherence', since in an RST-style analysis of a text every discourse unit is connected by at least one rhetorical link to at least another discourse unit. But in fact, this is often 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 (Knott et al. ress). Recent work on rhetorical relations is coming to the symmetrical position to ours–that

[52] Of course, identifying utterances with sentences ensures the presence of a link by merging together several clauses.

a purely relational account is not sufficient, and a separate theory of entity coherence is necessary (Knott et al. ress).[53]

Topic continuity: Rule 2 The other claim about entity coherence, Rule 2–stating a preference not just to keep talking about the same objects, but to preserve their relative ranking–was not tested in a statistical sense but seems much less robust, irrespective of its formulation.

The first point of interest about this aspect of the theory concerns the notion of transition that underlies it. With pretty much all parameter configurations that we tested, two of the most common transitions (if not the two most common) were two transitions not considered in the literature: the NULL transition between two utterances neither of which has a CB, and the ZERO transition from an utterance with a CB to one without. The question to be addressed 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).

The version of Rule 2 formulated in terms of sequences, and stating a preference for sequences of CON over sequences of RET over sequences of SHIFT, suffers from the problem that even in the 'best' versions more than two-thirds of sequences involve two different transitions. E.g., in the variant which yields the best results as far as Rule 2 is concerned, IF with second person pronouns and using Strube and Hahn's ranking, we find 143 CON-CON / RET-RET / SH-SH sequences, 47 EST-CON, 54 RET-SHIFT, 90 NULL-NULL, and 608 sequences of other types. Keeping this problem in mind, we do find in this version that the number of CON-CON sequences exceeds the number of RET-RET, which in turn exceeds the number of SH-SH. This doesn't hold with GFTHERELIN ranking, where RET-RET exceeds CON-CON unless we count EST-CON; nor for any of the other versions.

The formulations of Rule 2 based on single transitions, such as Brennan *et al.*, account for larger percentages of the total. However, we noted that a transition not discussed in these proposals, the NULL transition, is the most common transition in all but a few configurations (such as IF with GFTHERELIN, where the most common transition is EST). Also, that there are more RSH than SSH in most versions in which utterances are identified with sentences or allowing with indirect realization, the only exception being IF using the ranking proposed by Strube and Hahn. Finally, the preference for CON very much depends on whether we classify establishments as CON or not. If we do, CON is the most frequent transition in all of the 'best' versions. However, we observed when discussing in Section §5 our results concerning the linguistic predictions based on transitions that these correlations only hold, or are much stronger, if EST and CON are not conflated. For example, we saw 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.

As for the other versions of the Rule, we saw that Strube and Hahn's preference for sequences of cheap transitions over sequences over expensive ones isn't verified by any of the configurations we tested; indeed, in all configurations we looked at we found more expensive than cheap transitions.

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 (2.62) with indirect realization, u=s, treating second person entities as CFs, and Strube and Hahn ranking. In this version, more than 2 / 3 of utterances are continuous; on the other hand, only slightly more than 2 / 7 are cheap, salient or cohesive.[54]

[53]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.

[54]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.

Theoretical Proposals

To propose a new version of Centering is beyond the scope of this paper, but there are three broad theoretical conclusions that are suggested from these results, and should be further examined with psychological techniques.

The first conclusion is that entity-based accounts of coherence have to be integrated with accounts of other coherence-inducing factors. This could be done in two ways. We could be more explicit about the scope of Centering Theory, and view it not as a comprehensive account of 'local coherence', but of the contribution of entity coherence to local coherence. Alternatively, we could give a 'decomposed' formulation of Constraint 1, a bit like Kibble proposed for Rule 2. That is, we could list the factors that can link an utterance to the context, and propose that in order for an utterance to be 'locally coherent', at least one of these links must exist.[55]

The second conclusion is that perhaps Grosz and Sidner's idea that a single notion–the CB–is sufficient to account for both local salience and coherence is only an approximation. We may need separate conceptual tools: say, a CENTER OF COHERENCE to formulate Constraint 1 (and perhaps Rule 2) and a CENTER OF SALIENCE for Rule 1. The two centers might and often will be identical, but not at all times. In other words, it may be a good idea to reconsider Sidner's 'two foci' idea.

The third conclusion is that ensuring VARIETY seems to be as important a principle in discourse production as maintaining coherence. This is suggested by the fact that CBs are hardly ever continued for more than 2-3 utterances; that it is very unusual for the same discourse entity to be realized by the same type of NP twice in a row (even with pronouns, we only have 58 pronoun-pronoun sequences - 26% of the total); and that 2/3 of all transition sequences involve two different transitions. In fact, we hypothesize that the Repeated Name Penalty observed by Gordon *et al.*– 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–is but an instance of this more general phenomenon.

Definitional Issues raised by this study

Our experiments raised a number of questions about the definitions of the concepts used in Centering that we did not find mentioned in the literature, or were only discussed in passing.

A very important problem is the need to provide a definition of the notion of R1-pronoun: which anaphoric expressions are meant to be governed by Rule 1. Our provisional suggestion is that implicit anaphors such as traces should not be included; nor should second person pronouns. A second question is whether second person entities should be treated as CFs. The third question affects the theories in which ranking is based on grammatical function, and concerns the exact specification of grammatical function beyond the simplest cases. For example, should postcopular NPs in *there*-clauses be treated as subjects or objects? (Our results suggest the former.) And, how should nominal modifiers be ranked? (We treated them as adjuncts.) Finally, there is the question of how to determine the previous utterance when the 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 *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.*

[55] This would be a pretty weak formulation. We may propose a stronger one, requiring the existence of a 'strong link' - either an explicit anaphoric reference with a limited range of relations, or an explicit rhetorical connective.

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 'non-naturalistic' genres. It would be useful to include in the corpus texts from the genres more typically studied in Centering Theory, 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

We received lots of input on this work. Special thanks to Nikiforos Karamanis, Alistair Knott, Mark Liberman, and Ruslan Mitkov; and to 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, 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. 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. Janet Hitzeman was in part supported by the EPSRC project SOLE.

References

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

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.

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.

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 *roc. 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.

Carter, D. M. (1987). *Interpreting Anaphors in Natural Language Texts.* Ellis Horwood, Chichester, UK.

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.

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.

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., 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.

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. (The paper originally appeared as an unpublished manuscript in 1986.).

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.

Hardt, D. (1997). An empirical approach to VP ellipsis. *Computational Linguistics*, 23(4):525–541.

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. (1978). Resolving pronoun references. *Lingua*, 44:311–338.

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

Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385.

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.

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.

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

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

Keenan, E. L. and Westerståhl, D. (1997). Generalized quantifiers in linguistics and logic. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, pages 837–893. The MIT Press.

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

Kibble, R. (To appear). A reformulation of rule 2 of centering theory. *Computational Linguistics.*

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. (in press). 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.

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

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, USC, Information Sciences Institute.

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*.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proc. of the 18th COLING*, pages 869–875, Montreal.

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

Moore, J. D. (1995). *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context.* MIT Press, Cambridge, MA.

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.

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

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

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.

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. (2001a). 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.

Pearson, J., Stevenson, R., and Poesio, M. (2001b). Thematic roles, naming, and topicalisation as factors affecting the accessibility of the referent of a pronoun. Submitted.

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., 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., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA. AAAI.

Poesio, M. and Stevenson, R. (To appear). *Salience: Computational Models and Psychological 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.

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.

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.

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. and Hahn, U. (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. (1999). Analysis of syntax-based pronoun resolution methods. In *Proc. of the 37th ACL*, pages 602–605, University of Maryland. ACL.

Turan, U. (1995). *Null vs. Overt Subjects in Turkish Discourse: A Centering Approach*. PhD thesis, University of Pennsylvania. Also available as IRCS Report IRCS-96-13.

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.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4).

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

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.

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