J Log Lang Inf (2007) 16:445–464
DOI 10.1007/s10849-007-9048-z

Supplementing Entity Coherence with Local Rhetorical Relations for Information Ordering

Nikiforos Karamanis

Received: 19 July 2007 / Accepted: 19 July 2007 / Published online: 18 August 2007

Abstract This paper investigates whether the model of local rhetorical coherence suggested in Knott et al. (2001) can boost the performance of the Centering-based metrics of entity coherence employed by Karamanis et al. (2004) for the task of information ordering. Rhetorical coherence is integrated into the way Centering's basic data structures are derived from the annotated features of the GNOME corpus. The results indicate that (a) the simplest metric continues to perform better than its competitors even when local rhetorical coherence is taken into account, and (b) this extra coherence constraint decreases its performance.

Keywords Information ordering · Centering theory · Rhetorical coherence

1 Introduction

Text generation is the field in Computational Linguistics which deals with the automated production of text from information derived from either an underlying non-linguistic representation (concept-to-text generation: Reiter and Dale 2000) or other documents (text-to-text generation), e.g., to summarise them (Mani 2001). *Information Ordering* (Barzilay and Lee 2004), i.e., deciding in which sequence to present a set of preselected information-bearing items (typically corresponding to clauses or sentences) has received much attention in recent work in text generation. Text generation systems need to organise the content in a way that makes the output text *coherent*, i.e., easy to read and comprehend. The easiest way to exemplify coherence is by arbitrarily reordering the sentences of an understandable text. This process very often gives rise

N. Karamanis (✉)
Natural Language and Information Processing Group, Computer Laboratory,
University of Cambridge, Cambridge, UK
e-mail: Nikiforos.Karamanis@cl.cam.ac.uk

to documents that do not make sense although the information content is the same before and after the reordering (Marcu 1997; Reiter and Dale 2000).

Entity coherence, which is based on the way NP referents relate subsequent clauses in the text, is an important aspect of textual felicity. Since the early ’80s, when it was first introduced, *Centering Theory* has been an influential framework for modelling entity coherence, especially for text interpretation (see the collection of papers in Walker et al. 1998b for an overview). However, as Kibble (2001) observes, Centering began being applied to text generation only relatively recently.

Karamanis et al. (2004) presented the first attempt evaluate Centering-based metrics of coherence for the purposes of information ordering in text generation. A subset of the GNOME corpus (Poesio et al. 2004) was used as test data because it consisted of texts which were representative of the domain Karamanis et al. were mainly interested in, namely descriptions of museum artefacts, and were reliably annotated with features related to Centering. Although Centering was expected to be particularly appropriate for information ordering in this genre, their main finding was that the simplest metric sets a baseline that cannot be outperformed by other metrics which utilise additional Centering-specific notions. However, the baseline did not perform well enough to be used in practice for information ordering on its own.

Karamanis et al. tested metrics suitable for the information ordering approaches in text generation presented by Karamanis and Manurung (2002) and Althaus et al. (2004). These approaches, which are inspired by related work on text-to-text generation (Lapata 2003; Barzilay and Lee 2004; Barzilay and Lapata 2005), receive an unordered set of clauses as their input and use a metric to output the highest scoring ordering of these clauses.[1] The metrics were evaluated empirically using the experimental methodology of Karamanis (2003). The main assumption behind this method is that the observed ordering of clauses in a text represents a gold standard solution. The gold standard is scored by each metric, which is penalised proportionally to the amount of alternative orderings of the same material that score equally to or better than the gold standard. This methodology extends the way Barzilay and her colleagues evaluate automatically their approaches to information ordering.

Similarly to most work on Centering for text interpretation, Karamanis et al. investigated the impact of Centering only and did not take other coherence-inducing factors into account in their study. However, Kibble (2001) argued that Centering needs to be supplemented with other models of coherence while Poesio et al. (2004) suggested that the model of local rhetorical coherence introduced by Knott et al. (2001) may be a good candidate to supplement Centering in my domain of interest (i.e., object descriptions).

Knott et al. (2001) object to the traditional view of textual structure as a tree of Rhetorical Relations (RR-tree) motivated by Rhetorical Structure Theory (Mann and Thompson 1987). Organising the entire text structure hierarchically in terms of an RR-tree can be traced back to at least Hovy (1988) and has inspired a lot of work in text generation, with the approaches of Scott and de Souza (1990) and

[1] Typically, information ordering in concept-to-text generation is a side effect of building a tree of Rhetorical Relations. However, this is not the most appropriate way to account for the coherence of descriptive texts as I discuss in more detail below.

Marcu (1997) being among the most influential. Nevertheless, Knott et al. argue that descriptive texts do not feature an entirely tree-like structure. In their model of *local rhetorical coherence*, RR-trees are made of a small number of Rhetorical Relations applied locally. The local RR-trees are related to each other via links induced by constraints on entity coherence.

Knott et al. do not commit to a specific framework of entity coherence to supplement local rhetorical coherence in their model although Centering is mentioned as a potentially compatible theory. One way of integrating Centering with local rhetorical structure has been suggested by Kibble and Power (2000, 2004). In Kibble and Power's system, which generates pharmaceutical leaflets, notions derived from Centering are applied together with constraints on rhetorical coherence to decide on the best local RR-tree. Crucially, in Kibble and Power's approach Centering applies within portions of the local RR-tree. By contrast, in Knott et al.'s model the local RR-trees make up the units to which entity coherence applies.

This paper builds on the work of Karamanis et al. by supplementing entity coherence with local rhetorical relations to investigate whether the latter type of coherence can boost the results reported there. Given that my genre of interest is the same as Knott et al.'s, attending to their model appears appropriate. Since the exact nature of entity coherence in this model is underspecified, I define it more precisely in Centering terms as Kibble and Power did. Rhetorical coherence is then integrated into the way Centering's basic data structures are built, defining novel representations. Their novelty relies on the facts that: (i) Unlike Kibble and Power, entity coherence supersedes rhetorical constraints in my approach, thus staying closer to the spirit of Knott et al. for the purposes of concept-to-text generation. (ii) Text-to-text generation models such as those suggested by Barzilay and Lapata do not incorporate the notion of rhetorical coherence at all. The representations are also general in the sense that they are in principle applicable to both types of text generation, as explained in Sect. 2.2 in more detail.

These representations are used as the input to corpus-based experiments testing the effect of rhetorical coherence in the evaluation of Centering-based metrics for the first time. The results indicate that (a) the baseline remains the best performing metric when compared to its competitors even when local rhetorical coherence is taken into account, and (b) supplementing the baseline with this extra coherence constraint decreases its performance.

The paper is structured as follows: First, I outline Centering and explain how Centering data structures can be derived from the annotation features of the GNOME corpus (Sect. 2). Then, I discuss how local rhetorical coherence can be taken into account in this domain (Sect. 3). After a brief presentation of the Centering-based metrics of coherence and the experimental methodology (Sects. 4 and 5), the results of the study are presented (Sect. 6) and their implications are discussed (Sect. 7). The paper is concluded with an outline of related and future work (Sect. 8).

2 Centering Theory

How NP referents contribute to coherence is discussed in several seminal papers such as Chafe (1976), Kintsch and van Dijk (1978), Reinhart (1981), Givon (1983), and Horn (1986), among others. Centering was first introduced by Grosz et al. (1983) as a simple theory of entity coherence, attempting to model some aspects of immediate focus in the computational theory of Sidner (1979).[2] The theory was subsequently reformulated and an influential manuscript originating from the mid '80s was published in its final form as Grosz et al. (1995). This manuscript was circulated before its official publication and inspired a lot of work including the seminal paper of Brennan et al. (1987).

According to Grosz et al. (1995), *centers* (i.e., NP referents) are semantic objects that are part of the discourse model for each utterance and correspond to discourse *entities* in the sense of Webber (1978) or Kamp and Reyle (1993).[3]

Each utterance, U_n, is assigned a *list of forward-looking centers*, denoted as $\mathrm{CF}(\mathrm{U}_n)$, which represents a partial ranking of the NP referents in U_n in order of prominence. The *preferred center*, $\mathrm{CP}(\mathrm{U}_n)$, is the most highly ranked member of $\mathrm{CF}(\mathrm{U}_n)$. According to Centering's Constraint 3 (see below), the most highly ranked element of $\mathrm{CF}(\mathrm{U}_{n-1})$ that is realised in U_n is the *backward-looking center* $\mathrm{CB}(\mathrm{U}_n)$. The $\mathrm{CB}(\mathrm{U}_n)$ links the current utterance to the previous one. Obviously, segment-initial utterances lack a CB.

Note that $\mathrm{CB}(\mathrm{U}_n)$ can only come from $\mathrm{CF}(\mathrm{U}_{n-1})$ and not from prior sets of forward-looking centers. Additionally, Grosz et al. (1995) emphasise that there cannot be more than one CB for each utterance, a principle known as Constraint 1 (see below). As Kibble (2001) and Poesio et al. (2004) notice, the core proponents of Centering do not appear to explicitly acknowledge any other factor as being relevant to discourse coherence.

Based on the distinction between the CB and the CP, Brennan et al. (1987) defined four transition relations across pairs of adjacent utterances which, despite several variations,[4] are the most commonly used in the literature. Table 1 presents the four standard Centering transitions (Walker et al., 1998a, p. 6), the typology of which is based on two factors: whether the CB is the same from U_{n-1} to U_n, and whether the $\mathrm{CB}(\mathrm{U}_n)$ is the same as the $\mathrm{CP}(\mathrm{U}_n)$.[5]

The most popular versions of Centering (that I will subsequently refer to as "standard Centering") also make use of the following formal system of constraints and rules (Walker et al. 1998a, pp. 3–4):

[2] More details on how Centering relates with Sidner's model and the aforementioned theories are given in Miltsakaki (2003, Chap. 2).

[3] Strictly speaking, Centering was suggested as a model of local entity coherence, applying within discourse segments, to supplement the model of global entity coherence in Grosz and Sidner (1986). However, related work failed to identify discourse segments reliably (Passoneau 1998b) so Centering is typically applied throughout the whole text.

[4] The most important variation comes from Grosz et al. (1995) who define only one SHIFT transition using only the condition $\mathrm{CB}(\mathrm{U}_n) \neq \mathrm{CB}(\mathrm{U}_{n-1})$. Strube and Hahn (1999, Table 20, p. 333) define six transitions although the most important concept in their framework is the principle of CHEAPNESS(see Sect. 2.1.3).

[5] "$\mathrm{CB}(\mathrm{U}_{n-1})$ undef." in Table 1 stands for the cases where U_{n-1} does not have a CB (also see Sect. 2.1).

Table 1 Standard Centering transitions are defined according to whether the backward looking center, CB, is the same in two subsequent utterances, U_{n-1} and U_n, and whether the CB of the current utterance, $CB(U_n)$, is the same as its preferred center, $CP(U_n)$

		COHERENCE: $CB(U_n) = CB(U_{n-1})$ or $CB(U_{n-1})$ undef.	COHERENCE∗: $CB(U_n) \neq CB(U_{n-1})$
SALIENCE:	$CB(U_n) = CP(U_n)$	CONTINUE	SMOOTH-SHIFT
SALIENCE∗:	$CB(U_n) \neq CP(U_n)$	RETAIN	ROUGH-SHIFT

These identity checks are also known as the principles of COHERENCE and SALIENCE (Sect. 2.1.2), the violations of which are denoted with an asterisk in the table

For each utterance U_n:

Constraint 1. There is precisely one $CB(U_n)$.
Constraint 2. Every element of $CF(U_n)$ must be realised in U_n.
Constraint 3. The $CB(U_n)$ is the highest-ranked element of $CF(U_{n-1})$ realised in U_n.

Rule 1. If any element of $CF(U_{n-1})$ is realised by a pronoun in U_n, then the $CB(U_n)$ must be realised by a pronoun also.
Rule 2. Transition states are ordered. CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT:

CONTINUE >> RETAIN >> SMOOTH-SHIFT >> ROUGH-SHIFT

Constraints 1 and 3 have already been discussed and will form the main premises of my formulation of Centering. Constraint 2 is more open to interpretation and in this work it is taken to correspond to what Grosz et al. call "direct realisation" which means that only NPs explicitly mentioned in an utterance are allowed to introduce referents to the CF list.[6] Rule 1 is not discussed at all in this paper as pronominalisation is not related to this work. However, I make use of the preferences between transitions as specified by Rule 2.

2.1 Centering Variations

As Poesio et al. (2004) observe, several researchers developed various different versions of Centering. In this section, I review a recent analysis of Centering into the *prerequisite* of CONTINUITY and three underlying *principles*, namely COHERENCE, SALIENCE and CHEAPNESS. This analysis was claimed to further simplify Centering.

2.1.1 Continuity

Constraint 1 of standard Centering can be taken to presuppose that each utterance in the discourse refers to at least one entity in the utterance that precedes it (Poesio et al.

[6] Hence, direct realisation ignores bridging relations between referents (Clark 1977) for the computation of the CF list.

2004). Arguably, this requirement can be seen as a *prerequisite* for the computation of the standard Centering transitions in Table 1. The definition of the prerequisite of CONTINUITY in terms of Centering is as follows (Karamanis and Manurung 2002):

$$\text{CONTINUITY} : \mathrm{CF}(\mathrm{U}_{n-1}) \cap \mathrm{CF}(\mathrm{U}_n) \neq \emptyset$$

As previously mentioned, Grosz et al. (1995) do not discuss the effects of violations of Constraint 1 in the coherence of discourse. Kibble and Power (2000, Fig. 1) define the additional transition NOCB for the second member of a pair of utterances that do not have any entity in common, suggesting that a NOCB can be considered as the transition which causes the highest degradation of entity coherence.[7]

As shown in Table 1, the inverse case, i.e., when U_n has a CB but U_{n-1} does not have one, is classified as a CONTINUE or a RETAIN by Walker et al. (1998a). The additional transition ESTABLISHMENT is also often used to refer to such an utterance (Kameyama 1998b; Poesio et al. 2004).

2.1.2 Coherence and Salience

As Table 1 shows, the standard Centering transitions can be rephrased in terms of two general *principles* (Kibble 2001; Beaver 2004):[8]

$$\text{COHERENCE} : \mathrm{CB}(\mathrm{U}_n) = \mathrm{CB}(\mathrm{U}_{n-1})$$
$$\text{SALIENCE} : \mathrm{CB}(\mathrm{U}_n) = \mathrm{CP}(\mathrm{U}_n)$$

These principles, and their violations, denoted as COHERENCE∗ and SALIENCE∗, are only considered to arise when U_n, i.e., the second utterance in a pair, has a CB.

Kibble and Beaver notice that ranking COHERENCE over SALIENCE (denoted as COHERENCE>>SALIENCE) is a simpler way of stating the preferences over transitions in Rule 2. This is evident from Centering's declared preference for a RETAIN over a SMOOTH-SHIFT. Since a RETAIN only violates COHERENCE and a SMOOTH-SHIFT only violates SALIENCE, the preference for a RETAIN over a SMOOTH-SHIFT is an indirect way of stating that violating COHERENCE is more serious than violating SALIENCE. More generally, reformulating the preferences of Rule 2 directly in terms of the underlying principles instead of the set of transitions is argued to make the Centering model simpler and more transparent (Beaver 2004).

2.1.3 Cheapness

Another reformulation of Centering, named Functional Centering, is defined in Strube and Hahn (1999). Strube and Hahn introduce the principle of CHEAPNESS in order to

[7] Different types of NOCBtransitions are introduced by Passoneau (1998b), Di Eugenio (1998b) and Poesio et al. (2004), among others. Other researchers, however, consider the NOCB transition to be a type of ROUGH-SHIFT (Miltsakaki and Kukich 2004).

[8] Kibble (2001) uses the term COHESION instead of COHERENCE for the first of these principles. The terms used by Beaver (2004) are COHERE and ALIGN for COHERENCEand SALIENCE, respectively.

improve the way that standard Centering resolves certain cases of pronoun anaphora:

$$\textsc{cheapness} : \mathrm{CB}(\mathrm{U}_n) = \mathrm{CP}(\mathrm{U}_{n-1})$$

Strube and Hahn (1999, Table 21, p. 333) introduce a table of 36 transition pairs, labelled as "cheap", "expensive", or "not applicable" depending on whether CHEAPNESS holds for the second transition in the pair. Then, they redefine Rule 2 which now favours cheap transition pairs over expensive ones. This means that CHEAPNESS is given total priority over the other two underlying principles of Centering in their model.

In the next section, I present an example text from my experimental domain, namely the GNOME corpus, and explain how Centering's data structures can be derived from its annotation features.

2.2 Centering Data Structures in GNOME

GNOME-LAB is a subset of the GNOME corpus consisting of 20 descriptions of museum artefacts (Karamanis et al. 2004). The following example is a characteristic text from this subcorpus:

> (a) [Item 144]$_S$ is a torc. (b) [Its present arrangement]$_S$, twisted into three rings, may be a modern alteration; (c) [it]$_S$ should probably be a single ring, worn around the neck. (d) [The terminals]$_S$ are in the form of goats' heads. (1)

The text spans with indexes (a) to (d) correspond to annotated *finite units* in GNOME. Finite units were chosen by Karamanis et al. among other possibilities such as sentences to form the basis for their representations because they are very commonly used in Centering and may be deployed in text-to-text generation. Additionally, they seemed to correspond better to "facts" typically employed in concept-to-text generation systems such as MPIRO (Isard et al. 2003), which was the application more closely related to this investigation.[9] In this way, a representation was built which is potentially applicable for both text-to-text and concept-to-text generation.

Karamanis et al. used the computational tools of Poesio et al. (2004) to automatically derive the CF list for each finite unit from GNOME's annotation. Referents of NPs such as `de374` (that is, the referent of "Item 144" which can be taken to accord to the argument of a fact in a concept-to-text generation scenario similar to the one suggested by e.g. Kibble and Power) are ranked in each list according to their prominence (see Table 2).

More specifically and following Brennan et al. (1987), the referent of the NP which bears the grammatical role of the subject (indicated with the subscript S in the exam-

[9] Clearly, the deployed representations are not purely conceptual while realising them as surface text is not trivial either. However, the "facts" that typically serve as input for concept-to-text generation incorporate a lot of linguistic information too (Reiter and Dale 2000).

Table 2 The CP (i.e., first member of the CF list), the next referent, the CB, NOCBor standard Centering transition (Table 1) and violations of CHEAPNESS, SALIENCE and COHERENCE (denoted with an asterisk) for each unit in example (1) from the GNOME-LAB corpus

Unit	CF list: {CP,	next referent}	CB	Transition	
(1a)	{de374,	de375}	n.a.	n.a.	
(1b)	{de376,	de374, ... }	de374	RETAIN	
(1c)	{de374,	de379, ... }	de374	CONTINUE	
(1d)	{de380,	de381, ... }	–	NOCB	
			CHEAPNESS CB$_n$ =CP$_{n-1}$	SALIENCE CB$_n$ =CP$_n$	COHERENCE CB$_n$ =CB$_{n-1}$
(1a)	{de374,	de375}	n.a.	n.a.	n.a.
(1b)	{de376,	de374, ... }	√	*	√
(1c)	{de374,	de379, ... }	*	√	√
(1d)	{de380,	de381, ... }	n.a.	n.a.	n.a.

ple) is defined as the CP, i.e., the first member of the CF list. Referents with the same grammatical role are ranked according to the linear order of the corresponding NPs in the text (Sect. 3.1 provides an example of this). The derived sequence of CF lists is then used to compute other important Centering concepts:

- The CB, i.e., the referent that links the current CF list with the previous one such as de374 in unit (1b) of Table 2. This is defined according to Centering's Constraint 3 (see Sect. 2).
- NOCBs, that is, cases in which two subsequent CF lists do not have any referent in common as in unit (1d) of Table 2.[10]
- Standard Centering transitions (defined in Table 1), and the preferences between them as defined by Centering's Rule 2 (see Sect. 2). The transitions for example (1) are listed in Table 2.
- The principle of CHEAPNESSand the decomposition of Centering into the principles of SALIENCE and COHERENCE(discussed in Sect. 2.1): see Table 2 for examples.[11]

3 Local Rhetorical Coherence

Accounting for discourse coherence in terms of Rhetorical Relations is another popular approach in computational linguistics formulated in work such as Hobbs (1985) or, more recently, Asher and Lascarides (2003). However, the most popular framework,

[10] In order to stick with the assumption that referents correspond to arguments of facts typically used in concept-to-text generation, Karamanis et al. ignored the annotated bridging relation (Clark 1977) between the referent of "the terminals" de380 in (1d) and the referent of "it" de374 in (1c), by virtue of which de374 might be thought as being a member of the CF list of (1d).

[11] Notice that none of these constraints is applicable for units marked with the NOCB transition such as (1d).

especially within the concept-to-text generation community has been Rhetorical Structure Theory (RST). According to RST (Mann and Thompson 1987), a natural text can be described as a tree-like hierarchical structure with Rhetorical Relations applying recursively between adjacent spans of text as well as between larger text spans already related via a Rhetorical Relation.[12]

Although RST-based approaches to text structuring have been very popular within the generation literature (as exemplified by the seminal work of Hovy 1988; Scott and de Souza 1990; Marcu 1997), the appropriateness of this framework for certain genres has been challenged (see e.g., Kittredge et al. 1991; Power et al. 2003, among others). In this section, I review the analysis of Knott et al. (2001), which is most closely related to my genre of interest, namely descriptive texts.

As Knott et al. observe, in a standard RST analysis of a descriptive text, most of the material appears to be related via a specific kind of rhetorical relation called ELABORATION. In general, ELABORATION has been characterised as "the weakest of all rhetorical relations in that its semantic role is simply one of providing more detail" (Scott and de Souza 1990, p. 60). Knott et al. (2001) identified a number of additional general theoretical problems in the RST framework all related to ELABORATION and suggested that this relation be eliminated from the group of RST relations and replaced by a theory of entity coherence.

Moreover, Knott et al. argue that descriptive texts do not feature an entirely tree-like structure. In their model of *local rhetorical coherence*, Rhetorical Relations (minus ELABORATION) apply only locally. The local trees of Rhetorical Relations (RR-trees) are related to each other via links induced by constraints on entity coherence. The main operational unit in this framework is the *entity chain* which consists of a sequence of local RR-trees connected with each other linearly via subsequent entity links.

Most of the focus in Knott et al. is on arguing against the ELABORATION relation and in favour of the claim that other Rhetorical Relations apply only locally. Thus, as they also state in their conclusion, the structure within and between the entity chains is underspecified. Although Centering is mentioned as one of the models of entity coherence that can be possibly applied within this framework exactly how this can be done remains an open question. In the next section, I discuss how I further specified Knott et al.'s framework using the CF list, Centering's basic data structure, as the building block of the entity chains in my experimental domain.

3.1 Local RR-trees in GNOME

Since the GNOME corpus is not annotated for Rhetorical Relations, in my preliminary exploration of the texts in GNOME-LAB I looked for local RR-trees using a cue phrase

[12] Another way to account for text structure in concept-to-text generation is by using schemata (McKeown 1985). While schemata typically express frequently occurring, domain-dependent text structures that exhibit little variation, RST is an attempt to describe the structure of a wider variety of texts in terms of the combination of a more or less fixed set of rhetorical relations which are seen as the building elements from which all coherent texts are composed.

such as “because”, “but”, “so”, etc, as their signal.[13] 19 local RR-trees in 12 texts from GNOME-LAB were identified in that way.[14] These 12 texts form the subcorpus GNOME-RR. The remaining 8 texts are similar to example (1) in that they do not feature any signalled Rhetorical Relation and are taken to consist of units related to each other purely via entity links.

Example (2) features a typical local RR-tree in GNOME-RR:

(a) Access to the cartonnier’s lower half can only be gained by the doors at the sides, (b) because the table would have blocked the front. (2)

Similarly to Kibble and Power, I assume that RR-trees have already been formed prior to information ordering. However, instead of taking local RRs as consisting of CF lists (as they do), I remain closer to the framework of Knott et al. and defined a CF list for each local RR-tree. In this way, Centering can apply between CF lists of simple units (i.e. units not participating in a Rhetorical Relation) as well as CF lists of RR-trees (which consists of units connected via a Rhetorical Relation) as shown in Fig. 1. More specifically, in the representation of Karamanis et al. (2004), the units of example (2) give rise to two CF lists:

$$\begin{aligned}&\text{CF list of (2a)} : \{\mathtt{de12}, \mathtt{de13}, \ldots\}\\&\text{CF list of (2b)} : \{\mathtt{de9}, \mathtt{de18}\}\end{aligned} \tag{3}$$

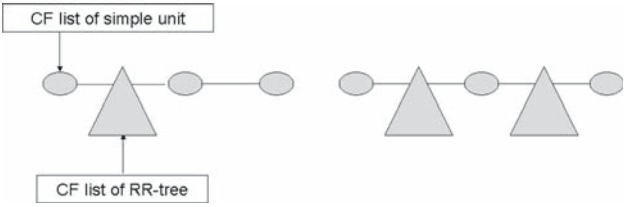

Fig. 1 Applying Centering between CF lists made of simple units and CF lists corresponding to local RR-trees

[13] Although cue phrases may not be the only hint for a Rhetorical Relation, it has been shown that they constitute a very reliable way of detecting one (Knott and Dale 1994) even when just one annotator is involved (which is the case here). On the other hand, absence of a cue phrase makes the detection of a relation particularly subjective. Thus, my assumption was that unsignalled relations (mainly) correspond to ELABORATIONS, which is clearly backed up by the corpus analyses of Marcu (1997) and Taboada (2006). Practically, this means that I take every unsignalled relation to be substitutable by entity coherence, which deviates slightly from Knott et al.’s model.

[14] Another concept in the framework of Knott et al., which is called *resumption*, accounts for relations between referents in non-subsequent units. Since I did not come across any examples of resumption in GNOME-LAB, this phenomenon was not taken into account.

The RR-tree in example (2) corresponds to an annotated *sentence* in GNOME.[15] Hence, the CF list of the RR-tree can be readily derived from the annotated data using *sentence* instead of *finite unit* as the basis for its computation (keeping the other Centering parameters such as the ranking of referents the same as in Karamanis et al. 2004). The first two members of the CF list for the sentence that contains (2a) and (2b) are shown in example (4):

Access to the cartonnier's lower half can only be gained by the
doors at the sides, because the table would have blocked the front. (4)
CF list of(4) : {`de12`, `de9`, ...}

The CF list of (4) replaces the CF lists of (2a) and (2b) in the data structures which are defined for the texts in GNOME-RR.

The sentence in (4) contains two NPs annotated as subjects: "Access to the cartonnier's lower half" (whose referent is `de12`) and "the table" (whose referent is `de9`). As mentioned in Sect. 2.2, referents with the same grammatical role are ranked according to the linear order of the corresponding NPs in the text. Thus, the CP of (4) is `de12` because "Access to the cartonnier's lower half" precedes "the table" within the sentence. If (2b) preceded (2a) within the sentence, the CP of (4) would have been `de9`.

In all but one case, the finite units that are related with each other via a Rhetorical Relation appear within the same sentence. The sentence consists only of these units, as in example (2), and the CF list can be computed automatically. The CF list of the sole RR-tree consisting of two finite units each forming a single sentence was computed by hand, using the surface order of the sentences for the ranking of referents with the same grammatical role.

Despite the isomorphism between RR-trees and sentences in GNOME-RR, it would be a mistake to consider the relationship between sentences consisting of more than one finite unit and RR-trees as 1:1. We identified 15 sentences in GNOME-LAB consisting of more than one finite unit which are not related to each other via an explicit Rhetorical Relation marked with a connective although they appear within the same sentence (units (1b) and (1c) are one such case).[16] These units are represented as subsequent, rhetorically unrelated, CF lists.[17]

Note that taking local RR-trees into account in this way reduces the overall number of CF lists a text is analysed to (and the corresponding number of possible orderings). More specifically, the texts in GNOME-RR contain 1.58 fewer CF lists when compared to the average number of CF lists in GNOME-LAB (8.35). About 23% of the CF lists in GNOME-RR lists are attributable to RR-trees.

[15] A sentence is defined as a span of text ending with a full stop, a question mark or an exclamation point (Poesio et al. 2004).

[16] While RR-trees typically correspond to subordinated finite units within the sentence, these are mainly examples of coordination between units. Taboada (2006) makes a similar observation.

[17] A more detailed and general study on the lack of isomorphism between document and rhetorical structure, motivated mainly by examples from GNOME's pharmaceutical section appears in Power et al. (2003).

4 Metrics of Coherence

The information ordering approaches in Karamanis and Manurung (2002) and Althaus et al. (2004) receive an unordered set of clauses as their input and use a metric of coherence to output the highest scoring ordering of these clauses. Clauses are represented in terms of CF lists and coherence is estimated using features from Centering. For each candidate ordering (defined as a permutation of the input set of CF lists), the metric computes a coherence score which is used to compare the ordering with its alternatives.

To make this clearer, let us first assume that the ordering of CF lists in Table 2 is a candidate ordering. The other possible ways of ordering those CF lists are its alternatives.

The candidate ordering of Table 2 contains one NOCB (unit 1d) so its score according to M.NOCB, the simplest metric employed in Karamanis et al. (2004), is 1. An alternative ordering without any NOCBs (provided that such an ordering exists) will be preferred over this ordering as the selected output of information ordering if M.NOCB is used to guide the process.[18] M.NOCB was used as the baseline in the experiments of Karamanis et al. because it does away with all Centering concepts expect for the prerequisite of CONTINUITY (Sect. 2.1.1).

M.CHEAP is based on the formulation of Functional Centering (Sect. 2.1.3) and gives preference to orderings with the fewest violations of CHEAPNESS.[19] The only unit which violates CHEAPNESS in Table 2 is (1c) so the score of the candidate ordering according to M.CHEAP is 1. Should an alternative ordering with fewer violations of CHEAPNESS exist, it will be used as the output according to M.CHEAP.

M.KP, introduced by Kibble and Power (2000), sums up the NOCBs as well as the violations of CHEAPNESS, COHERENCE and SALIENCE, preferring the ordering with the lowest total cost.[20] In addition to the aforementioned violations, the candidate ordering of Table 2 also violates SALIENCE once (unit 1c). Hence, its score according to M.KP is 3 and an alternative ordering with a lower score (if any) will be preferred by this metric.

Metrics such as M.KP are agnostic with respect to the occurrence of violations of different underlying principles in the same utterance. By contrast, the formulations of Centering which use transitions define a vocabulary for the way that such violations are combined in the same utterance.

M.BFP employs the transition preferences of Rule 2 (Sect. 2) and rewards the orderings which *maximise* preferred transitions such as CONTINUE, instead of those that *minimise* violations. The first score to be computed by M.BFP is the sum of CONTINUE transitions, which is 1 for the candidate ordering of Table 2 (due to unit 1c). This is used to compare it with its alternatives and if the candidate ordering is found to score higher than all of them it is selected as the output. If an alternative ordering is found

[18] If the best coherence score is assigned to an alternative ordering as well as the candidate ordering, then the information ordering algorithm will choose randomly between them.

[19] For CHEAPNESSto apply, CONTINUITY needs to have been satisfied first. This is why M.NOCB is taken as the baseline instead of M.CHEAP.

[20] A more recent variant of this metric appears in Kibble and Power (2004).

to have the same number of CONTINUES, the sum of RETAINS is examined. The sum of SMOOTH-SHIFTS is examined only when the orderings are found to have the same scores for the two more highly ranked transitions, etc.

Note that, following Brennan et al. (1987), NOCBS are not taken into account for the definition of transitions in M.BFP. Assume that the number of NOCB transitions for ordering T_1 is t_1 and the number of CONTINUES is c_1. In addition, ordering T_2 has t_2 NOCBS and c_2 CONTINUES. If both $t_1 > t_2$ and $c_1 > c_2$ hold, T_1 will loose the competition with T_2 according to M.NOCB but win it according to M.BFP.

Crucially, these are not the only ways to to define Centering-based metrics of coherence. Several ways of *ranking* the underlying principles of Centering have been suggested in the literature, most recently by Kibble (2001), Beaver (2004) and Kibble and Power (2004). Alongside Kibble (2001), I believe that the exact ranking of these concepts remains an open question.

Another possibility is to investigate transition-based metrics other than M.BFP. For instance, one may experiment with the definition of SHIFT in Grosz et al. (1995), the transitions enumerated in Strube and Hahn (1999, Table 20, p. 333), as well as the different versions of NOCB and ESTABLISHMENT mentioned in Sect. 2.1.1. These and numerous other possibilities are discussed in more detail in Karamanis (2003, Chap. 3), which also provides a formal definition of the metrics discussed above. In this paper, I abide by the metrics that outlined in Karamanis et al. (2004), aiming to investigate whether taking Rhetorical Relations into account will boost their performance.

5 Experimental Methodology

Since investigating coherence effects on naturally occurring discourse through psycholinguistic studies is almost infeasible, computational corpus-based experiments are the most viable alternative (Poesio et al. 2004; Barzilay and Lee 2004). This type of study is also particularly appropriate for problems which may involve a large number of parameters such as the wide range of Centering-derived metrics. In this section, I outline the corpus-based experimental methodology of Karamanis (2003) as the empirical framework under which I attempt to resolve the competition between a potentially large number of metrics.

This evaluation methodology is based on the premise that the gold standard ordering (GSO) of the clauses (and the corresponding CF lists) observed in a text is more coherent than any other ordering. If a metric takes an alternative ordering to be more coherent than the GSO, it has to be penalised.

Karamanis (2003) introduced a measure called the *classification error rate* which estimates this penalty as the weighted sum of the percentage of alternative orderings that score equally to or better than the GSO. The classification error rate is computed according to the following formula:[21]

$$\text{Better(M,GSO)} + \text{Equal(M,GSO)}/2$$

[21] Equal(M,GSO) is weighted by 1/2 on the basis of the assumption that, similarly to tossing a coin, the GSO will on average do better than half of the orderings that score the same as it does when other coherence constraints are taken into account. See Karamanis (2003, Chap. 5) for more details.

Table 3 Comparing the baseline metric M.NOCB with M.CHEAP, M.KP and M.BFP in GNOME-RR (left) and GNOME-LAB (right). The baseline outperforms the three other metrics in both corpora

	GNOME-RR corpus M.NOCB				GNOME-LAB corpus M.NOCB			
	Lower	Greater	Ties	*p*	Lower	Greater	Ties	*p*
M.CHEAP	10	2	0	.038	18	2	0	.000
M.KP	11	1	0	.006	16	2	2	.002
M.BFP	7	5	0	.774	12	3	5	.036
N of texts	12				20			

Better(M,GSO) stands for the percentage of orderings that score better than the GSO according to a metric M, whilst Equal(M,GSO) is the percentage of orderings that score equal to the GSO. When comparing several metrics with each other, the one with the lowest classification error rate is the most appropriate for ordering the CF lists that the GSO consists of.

In this study, I use the classification error rate to measure the performance of the metrics and investigate the following questions: (a) Is the best performing metric in GNOME-RR different from the one in GNOME-LAB? (b) Does taking local RR-trees into account improve the performance of the metrics?

6 Results

6.1 Which is the Best Metric?

The experimental results of the comparisons of the metrics from Sect. 4 are reported in Table 3 (left). Following Karamanis et al. (2004), the tables compare the baseline metric M.NOCB with each of M.CHEAP, M.KP and M.BFP. The exact number of texts (GSOs) for which the classification error rate of M.NOCB is lower (i.e., better) than its competitor for each comparison is reported in the second column of the Table.

For example, M.NOCB has a lower classification error rate than M.CHEAP for 10 (out of 12) GSOs from GNOME-RR. M.CHEAP achieves a lower classification error rate for just 2 GSOs, while there are no ties, i.e., cases in which the classification error rate of the two metrics is the same. The *p* value returned by the two-tailed sign test for the difference in the number of GSOs, rounded to the third decimal place, is reported in the fifth column of Table 3.[22]

Overall, the Table shows that M.NOCB does significantly better than M.CHEAP and M.KP in GNOME-RR. Since M.BFP fails to significantly outperform M.NOCB, the baseline can be considered the most promising solution in that case too by applying Occam's razor. This in turn indicates that simply avoiding NOCB transitions is

[22] The sign test was chosen by Karamanis et al. (2004) over its parametric alternatives to test significance because it does not carry specific assumptions about population distributions and variance and is more appropriate for small sample sizes.

Table 4 Changes in the classification error rate of the metrics in GNOME-RR

	GNOME-RR corpus			
Metric	Lower	Greater	Ties	*p*
M.NOCB	3	9	0	.146
M.CHEAP	9	3	0	.146
M.KP	10	2	2	.038
M.BFP	5	7	5	.774
N of texts	12			

more relevant to information ordering than the various combinations of the Centering notions that the other metrics make use of.

The right section of Table 3 shows the results of the evaluation of the metrics in GNOME-LAB from Karamanis et al. (2004). These are very similar to the ones just reported in that the baseline significantly outperforms the three other metrics which employ additional Centering concepts. Hence, M.NOCB is the most suitable among the investigated metrics for information ordering in this domain irrespective of whether local RR-trees are taken into account for the computation of the CF list.

6.2 Does Rhetorical Coherence Help?

I was also interested to see whether taking RR-trees into account improves the performance of the metrics in absolute terms. To do this I compared the classification error rates of the 12 GSOs for each metric in GNOME-RR with the corresponding classification error rates in GNOME-LAB (Table 4). The Table suggests that using local RR-trees for the computation of the CF list lowers the classification error rate for most GSOs, i.e., improves the performance of M.CHEAP as well as M.KP (for which the difference is significant). However, since both these metrics are defeated overwhelmingly by M.NOCB (see previous section), this improvement seems to be of little use.

Notably, M.NOCB continues to beat its opponents despite the fact that its own classification error rates are increased (i.e., performance is worsened) in 9 out of 12 GSOs. This observation is coupled by the value of the *average classification error rate* (Karamanis 2003) of M.NOCB which is an estimate of how likely M.NOCB is to come up with the GSO if it is actually used to guide an algorithm which orders the CF lists in the corpora. The average classification error rate in GNOME-RR is 23.24%, which means that on average M.NOCB takes approximately 1 out of 4 alternative orderings in GNOME-RR to be more coherent than the GSO. This compares poorly with the value of 19.95% in GNOME-LAB and suggests that RRs do not help M.NOCB become more efficient for information ordering in the investigated domain. In the following section, I discuss the implications of the experimental results.

7 Discussion

The small size of the corpus deployed in this work may arguably serve as the main criticism against it. However, the texts I experimented with are considered to be representative of the descriptive genre in general (Poesio et al. 2004) and several effects *are* strong enough to reject the null hypothesis on the basis of statistical tests even in this small corpus. Although I acknowledge that this work does not contribute any new resources (which are clearly required for more extensive experimentation), it introduces novel representations which can be used as the basis for more extended investigations in the combinations of entity and rhetorical coherence in the future. Additionally and despite its limitations, my preliminary empirical study enables several interesting observations to be made.

As already pointed out in Karamanis et al. (2004), the results suggest that if one is provided with the set of CF lists from a GSO in the domain of interest and has to choose which of the four candidate metrics to use to order them (aiming to arrive at the GSO as the output), the baseline M.NOCB is a better choice than M.KP, M.CHEAP and M.BFP. This is because there exist proportionally fewer alternative orderings that are taken to be more coherent than the GSO according to M.NOCB in comparison to the coherence assessments made by the other metrics.

Avoiding NOCBs is hardly a Centering-specific requirement and is typically seen as just a prerequisite for computing other more Centering-related notions. This work shows that NOCBs are much more useful than several such notions as far as information ordering is concerned. Of course, other Centering concepts remain very important for tasks such as anaphora resolution (for which notions such as CHEAPNESS were originally introduced).

My empirical results indicate that the predominance of M.NOCB holds irrespective of whether local RR-trees are taken into account for the computation of the CF lists. Thus, M.NOCB is a very robust baseline against which other, perhaps even more informed, metrics may be compared.

Poesio et al. report that disfavoured transitions such as NOCBs are very frequent in GNOME, which leads them to the conclusion that Centering needs to be supplemented with another model of coherence such as the one suggested by Knott et al. Using a hybrid model of entity and rhetorical coherence is also adopted by text-to-text generation practitioners such as Kibble and Power.

The majority of transitions in both GNOME-LAB and GNOME-RR (57% and 53%, respectively) are NOCBs. This accords with the findings of Poesio et al. and might cause one to think that entity coherence has indeed little to do with the investigated domain.

However, using the classification error rate to estimate the effect of entity coherence for information ordering in this domain sheds new light into this issue. The average classification error rate of M.NOCB is approximately 20% in GNOME-LAB and 23% in GNOME-RR. This suggests that the GSO tends to be in greater agreement with the preference to avoid NOCBs that the overwhelming majority (i.e., 80% in GNOME-LAB and 77% in GNOME-RR) of alternative orderings. In this sense, it seems that the observed ordering in the corpus (that is, the GSO) does optimise with respect to the number of potential NOCBs to a great extent. This is not obvious if the effect of entity coherence is estimated simply on the basis of the transition frequencies as it has been done until now.

Since the number of possible orderings becomes smaller when local RR-trees are taken into account, the information ordering problem is somewhat simplified. One might also be tempted to think that since GNOME-RR contains 4% fewer NOCBtransitions than GNOME-LAB, computing the CF list using local RR-trees should be adopted for information ordering.

However, the 3% rise in the aforementioned classification error rates provides evidence that there exist proportionally more orderings which are taken to be more coherent than the GSO in GNOME-RR than in GNOME-LAB. Thus, taking local RR-trees into account does not help M.NOCB improve its performance. This in turn indicates that a solution based on the model of Knott et al. is not particularly helpful, at least as far as information ordering in this domain is concerned.

Overall, my empirical results clarify which aspects of entity and local rhetorical coherence are more relevant to information ordering and puts other related work into perspective. These experiments also provide researchers working in information ordering with a simple and easily extendable evaluation framework as well as a robust baseline to deploy for their own meaningful comparisons.

8 Related and Future Work

The automatic evaluation of information ordering has received considerable attention in recent years. Lapata (2006) is the latest contribution in this line of work, extending the work of Lapata (2003) with additional empirical results. Lapata's measure as well as the one used by Barzilay and Lee (2004) are discussed in comparison with the classification error rate in Karamanis and Mellish (2005). Karamanis (2003, Chap. 9) present an example of how several measures can be combined in experiments deployed on facts derived from the MPIRO concept-to-text generation system (Isard et al. 2003) and ordered by domain experts.

In other related work, I deployed the purely Centering-based metrics to several additional domains: (a) 122 orderings of MPIRO facts (a corpus introduced by Dimitromanolaki and Androutsopoulos 2003) and (b) 200 newspaper articles and 200 accident narratives collected by Barzilay and Lapata (2005). As reported in Karamanis (2003) and Karamanis (2006), the results from these domains verify the ones stated in Karamanis et al. (2004) with the baseline overwhelmingly beating its competitors. As the formulations deployed there do not take rhetorical coherence into account, extending this work to that direction is desirable, albeit non-trivial, and might shed some light on how domain specific the trends reported in this paper are.

The enrichment of the deployed metrics with additional constraints of coherence remains the biggest challenge for the work reported in this paper. Initial results indicate that making use of features related to global focus in GNOME-LAB has the same effect as local RR-trees, i.e., they increase—instead of reduce—the classification error rate of the metrics, as reported in Karamanis (2003, Chap. 8). These features were used to further extend the representations in GNOME-RR as well (aiming to investigate the interaction between global focus and rhetorical coherence), but the limited size of the corpus did not allow any significant observations to be made in this occasion. The extension of the GNOME corpus with additional texts from the descriptive genre would have been a particularly welcome development, which unfortunately I was not in a position to materialise during this project.

Given the abundance of possible Centering-based metrics and several different ways of instantiating Centering, one might be keen to investigate whether a different metric might outperform M.NOCB or whether using bridging or sentences for the computation of the CF list affects its performance (also when local RR-trees are used). This will be more straightforward to do using the extant data, although these representations will probably be less applicable to concept-to-text generation than the one I deployed.

Last but not least, the evaluation in this paper is based on purely corpus-based methods. These should ideally be supplemented with human judgments in the spirit of the work reported by Reiter and Sripada (2002), Barzilay and Lapata (2005) and Lapata (2006).

Acknowledgements I am grateful to Massimo Poesio, Chris Mellish and Jon Oberlander for their invaluable guidance and advice and to James Soutter for significant programming assistance. Also many thanks to the participants of the "Coherence in Generation Workshop" in ESSLLI 2006, where an earlier version of this paper was presented, for their feedback and to Ruth Seal and Ian Lewin and three anonymous reviewers for their comments. Support from the Greek State Scholarships Foundation (IKY) and the BBSRC-funded FlySlip project (grant No 38688) is also acknowledged.

References

Althaus, E., Karamanis, N., & Koller, A. (2004). Computing locally coherent discourses. In *Proceedings of ACL 2004* (pp. 399–406). Barcelona, Spain.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of ACL 2005* (pp. 141–148).

Barzilay, R., & Lee, L. (2004). Catching the drift: Probabilistic content models with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004* (pp. 113–120).

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

Brennan, S. E., Friedman [Walker], M. A., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of ACL 1987* (pp. 155–162). Stanford, California.

Chafe, W. (1976). Giveness, contrastiveness, definiteness, subjects, topics and point of view. In C. Li (Ed.), *Subject and Topic* (pp. 25–76). New York: Academic Press.

Clark, H. H. (1977). Bridging. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 9–27). Cambridge University Press.

Di Eugenio, B. (1998b). Centering in Italian. In M. A. Walker, A. K. Joshi & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 115–137). Oxford: Clarendon Press.

Dimitromanolaki, A., & Androutsopoulos, I. (2003). Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation* (pp. 23–30). Budapest, Hungary.

Givon, T. (Ed.). (1983). *Topic continuity in discourse: A quantitative cross-language study*. John Benjamins.

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

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of ACL 1983* (pp. 44–50). Cambridge, MA.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203–225.

Hobbs, J. (1985). On the coherence and structure of discourse. Research report 85–37, CSLI.

Horn, L. R. (1986). Presupposition, theme and variations. *Chicago Linguistic Society, 22*, 168–192.

Hovy, E. (1988). Planning coherent multisentential text. In *Proceedings of ACL 1988* (pp. 163–169).

Isard, A., Oberlander, J., Androutsopoulos, I., & Matheson, C. (2003). Speaking the users' languages. *IEEE Intelligent Systems Magazine, 18*(1), 40–45.

Kameyama, M. (1998b) Intrasentential centering: A case study. In M. A. Walker, A. K. Joshi & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 89–122). Oxford: Clarendon Press.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to Modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Dordrecht: Kluwer.

Karamanis, N., & Manurung, H. M. (2002). Stochastic text structuring using the principle of continuity. In *Proceedings of INLG 2002* (pp. 81–88). Harriman, NY, USA, July 2002.

Karamanis, N., & Mellish, C. (2005). A review of recent corpus-based methods for evaluating information ordering in text production. In *Proceedings of Corpus Linguistics 2005 Workshop on Using Corpora for NLG* (pp. 13–18).

Karamanis, N., Poesio, M., Mellish, C., & Oberlander, J. (2004). Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Proceedings of ACL 2004* (pp. 391–398). Barcelona, Spain.

Karamanis, N. (2003). Entity Coherence for Descriptive Text Structuring. PhD thesis, Division of Informatics, University of Edinburgh.

Karamanis, N. (2006) Evaluating centering for information ordering in two new domains. In *Proceedings of NAACL 2006*.

Kibble, R., & Power, R. (2000). An integrated framework for text planning and pronominalisation. In *Proceedings of INLG 2000* (pp. 77–84). Israel.

Kibble, R., & Power, R. (2004). Optimizing referential coherence in text generation. *Computational Linguistics, 30*(4), 401–416.

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

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

Kittredge, R., Korelsky, T., & Rambow, O. (1991). On the need for domain communication knowledge. *Computational Intelligence, 7*, 305–314.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes, 18*(1), 35–62.

Knott, A., Oberlander, J., O'Donnell, M., & Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (chap. 7, pp. 181–196). Amsterdam: John Benjamins.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003* (pp. 545–552). Saporo, Japan, July 2003.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's Tau. *Computational Linguistics, 32*(4), 1–14.

Mani, I. (2001). *Automatic summarization*. John Benjamins Publishing Co.

Mann, W. C., & Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organisation. Technical Report RR-87-190, University of Southern California/Information Sciences Institute.

Marcu, D. (1997). The Rhetorical Parsing, Summarisation and Generation of Natural Language Texts. PhD thesis, Department of Computer Science, University of Toronto.

McKeown, K. (1985). *Text generation: using discourse strategies and focus constraints to generate natural language Text*. Studies in Natural Language Processing. Cambridge University Press.

Miltsakaki, E., & Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Journal of Natural Language Engineering, 10*(1), 25–55.

Miltsakaki, E. (2003). The Syntax-Discourse Inteface: Effects of the Main-Subordinate Distinction on Attention Structure. PhD thesis, Department of Linguistics, University of Pennsylvania.

Passoneau, R. J. (1998b). Interaction of discourse structure with explicitness of discourse anaphoric phrases. In M. A. Walker, A. K. Joshi & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 327–358). Oxford: Clarendon Press.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics, 30*(3), 309–363.

Power, R., Scott, D., & Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics, 29*(2), 221–260.

Reinhart, T. (1981). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica, 27*, 53–94.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Reiter, E., & Sripada, S. (2002). Should corpora texts be gold standards for NLG? In *Proceedings of INLG 2002* (pp. 97–104). Harriman, NY, USA, July 2002.

Scott, D., & de Souza, C. S. (1990). Getting the message across in RST-based text generation. In R. Dale, C. Mellish, & M. Zock (Eds.), *Current research in natural language generation* (pp. 47–74). Academic Press.

Sidner, C. L. (1979). Towards a Computational Theory of Definite Anaphora Comprehension in English. PhD thesis, AI Laboratory/MIT, Cambridge, MA, June 1979. Also available as Technical Report No. AI-TR-537.

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

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics, 38*(4), 567–592.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998a). Centering in naturally occurring discourse: An overview. In Walker et al. (pp. 1–30).

Walker, M. A., Joshi, A. K., & Prince, E. F. (Eds.). (1998b). *Centering theory in discourse*. Oxford : Clarendon Press.

Webber, B. L. (1978). *A Formal Approach to Discourse Anaphora*. PhD thesis, Harvard University.