Evaluating Centering for Information Ordering using Corpora

Nikiforos Karamanis*
University of Cambridge

Chris Mellish**
University of Aberdeen

Massimo Poesio†
University of Essex

Jon Oberlander‡
University of Edinburgh

In this paper we discuss several metrics of coherence defined using Centering Theory and investigate the usefulness of such metrics for information ordering in automatic text generation. We estimate empirically which is the most promising metric and how useful this metric is using a general methodology applied on several corpora. Our main result is that the simplest metric (which relies exclusively on NOCB *transitions) sets a robust baseline that cannot be outperformed by other metrics which make use of additional Centering-based features. This baseline can be used for the development of both text-to-text and concept-to-text generation systems.*

1. Introduction

Information ordering (Barzilay and Lee 2004), that is, deciding in which sequence to present a set of preselected information-bearing items, has received much attention in recent work in automatic text generation. This is because text generation systems need to organise the content in a way that makes the output text *coherent*, i.e. easy to read and understand. The easiest way to exemplify coherence is by arbitrarily reordering the sentences of a comprehensible text. This process very often gives rise to documents that do not make sense although the information content is the same before and after the reordering (Hovy 1988; Marcu 1997; Reiter and Dale 2000).

Entity coherence, which is based on the way the referents of Noun Phrases (NPs) relate subsequent clauses in the text, is an important aspect of textual organisation. Since the early '80s, when it was first introduced, *Centering Theory* has been an influential framework for modelling entity coherence. Seminal papers on Centering such as Brennan, Friedman [Walker], and Pollard (1987, p.160) and Grosz, Joshi, and Weinstein (1995, p.215) suggest that Centering may provide solutions for information ordering.

Indeed, following the pioneering work of McKeown (1985), recent work on text generation exploits constraints on entity coherence to organise information (Mellish et al. 1998; O'Donnell et al. 2001; Cheng 2002; Kibble and Power 2000, 2004; Lapata 2003; Barzilay and Lee 2004; Barzilay and Lapata 2005, among others). Although these

* Computer Laboratory, William Gates Building, Cambridge CB3 0FD, UK. Nikiforos.Karamanis@cl.cam.ac.uk
** Department of Computing Science, King's College, Aberdeen AB24 3UE, UK.
† Department of Computer Science, Wivenhoe Park, Colchester CO4 3SQ, UK.
‡ School of Informatics, 2 Buccleuch Place, Edinburgh EH8 9LW, UK.

Submission received: 15 May 2006; revised submission received: 15 December 2007; accepted for publication: 7 January 2008.

approaches often make use of heuristics related to Centering, the features of entity coherence they employ are usually defined informally. Additionally, Centering-related features are combined with other coherence-inducing factors in ways that are based mainly on intuition, leaving many equally plausible options unexplored.

Thus, the answers to the following questions remain unclear: (i) *How appropriate is Centering for information ordering in text generation?* (ii) *Which aspects of Centering are most useful for this purpose?* These are the issues we investigate in this paper which presents the first systematic evaluation of Centering for information ordering. To do this, we define Centering-based *metrics of coherence* which are compatible with several extant information ordering approaches. An important insight of our work is that Centering can give rise to many such metrics of coherence. Hence, a general methodology for identifying which of these metrics represent the most promising candidates for information ordering is required.

We adopt a corpus-based approach to compare the metrics empirically and demonstrate the portability and generality of our evaluation methods by experimenting with several corpora. Our main result is that the simplest metric (which relies exclusively on NOCB transitions) sets a baseline that cannot be outperformed by other metrics which make use of additional Centering-related features. Thus, we provide substantial insight into the role of Centering as an information ordering constraint and offer researchers working on text generation a simple, yet robust, baseline to use against their own information ordering approaches during system development.

The paper is structured as follows: In section 2 we discuss our information ordering approach in relation to other work on text generation. After a brief introduction to Centering in section 3, section 4 demonstrates how we derived Centering data structures from existing corpora. Section 5 discusses how Centering can be used to define various metrics of coherence suitable for information ordering. Then, section 6 outlines a corpus-based methodology to choose among these metrics. Section 7 reports on the results of our experiments and section 8 discusses their implications. We conclude the paper with directions for future work and a summary of our main contributions.[1]

2. Information Ordering

Information ordering has been investigated by substantial recent work in text-to-text generation (Barzilay, Elhadad, and McKeown 2002; Lapata 2003; Barzilay and Lee 2004; Barzilay and Lapata 2005; Siddharthan 2006; Ji and Pulman 2006; Bollegala, Okazaki, and Ishizuka 2006; Soricut and Marcu 2006; Madnani et al. 2007, among others) as well as concept-to-text generation (particularly Kan and McKeown (2002) and Dimitromanolaki and Androutsopoulos 2003).[2] We added to this work by presenting approaches to information ordering based on a genetic algorithm (Karamanis and Manurung 2002) and linear programming (Althaus, Karamanis, and Koller 2004) which can be applied to both concept-to-text and text-to-text generation. These approaches use a metric of coherence defined using features derived from Centering and will serve as the premises of our investigation of Centering in this paper.

Metrics of coherence are used in other work on text generation too (Mellish et al. 1998; Cheng 2002; Kibble and Power 2000, 2004). With the exception of Kibble and

1 Earlier versions of this work were presented in Karamanis et al. (2004) and Karamanis (2006).
2 Concept-to-text generation is concerned with the automatic generation of text from some underlying non-linguistic representation. By contrast, the input to text-to-text generation applications is text.

Power's work, the features of entity coherence used in these metrics are informally defined using heuristics related to Centering. Additionally, the metrics are further specified by combining these features with other coherence-inducing factors such as rhetorical relations (Mann and Thompson 1987). However, as acknowledged in most of this work, these are preliminary computational investigations of the complex interactions between different types of coherence which leave many other equally plausible combinations unexplored.

Clearly, one would like to know what Centering can achieve on its own before devising more complicated metrics. To address this question, we define metrics which are *purely* Centering-based, placing any attempt to specify a more elaborate model of coherence beyond the scope of this paper. This strategy is similar to most work on Centering for text interpretation in which additional constraints on coherence are not taken into account (the papers in Walker, Joshi, and Prince (1998) are characteristic examples). This simplification makes it possible to assess for the first time how useful the employed Centering features are for information ordering.

Work on text generation which is solely based on rhetorical relations (Hovy 1988; Marcu 1997, among others) typically masks entity coherence under the ELABORATION relation. However, ELABORATION has been characterised as "the weakest of all rhetorical relations" (Scott and de Souza 1990, p.60). Knott et al. (2001) identified several theoretical problems all related to ELABORATION and suggested that this relation be replaced by a theory of entity coherence for text generation. Our work builds on this suggestion by investigating how appropriate Centering is as a theory of entity coherence for information ordering.

McKeown (1985, pp.60-75) also deployed features of entity coherence to organise information for text generation. McKeown's "constraints on immediate focus" (which are based on the model of entity coherence that was introduced by Sidner (1979) and precedes Centering) are embedded within the schema-driven approach to generation which is rather domain-specific (Reiter and Dale 2000). By contrast, our metrics are general and portable across domains and can be applied within information ordering approaches which are applicable to both concept-to-text and text-to-text generation.

3. Centering Overview

This section provides an overview of Centering, focusing on the aspects which are most closely related to our work. Poesio et al. (2004) and Walker, Joshi, and Prince (1998) discuss Centering and its relation to other theories of coherence in more detail.

According to Grosz, Joshi, and Weinstein (1995), each utterance U_n is assigned a ranked list of forward looking centers (i.e. discourse entities) denoted as CF(U_n). The members of CF(U_n) must be realised by the NPs in U_n (Brennan, Friedman [Walker], and Pollard 1987). The first member of CF(U_n) is called the preferred center CP(U_n).

The backward looking center CB(U_n) links U_n to the previous utterance U_{n-1}. CB(U_n) is defined as the most highly ranked member of CF(U_{n-1}) which also belongs to CF(U_n). CF lists prior to CF(U_{n-1}) are not taken into account for the computation of CB(U_n). The original formulations of Centering by Brennan, Friedman [Walker], and Pollard (1987) and Grosz, Joshi, and Weinstein (1995) lay emphasis on the uniqueness and the locality of the CB and will serve as the foundations of our work.

The CB and the CP are combined to define transitions across pairs of adjacent utterances (Table 1). This definition of transitions is based on Brennan, Friedman [Walker], and Pollard (1987) and has been popular with subsequent work.

Table 1
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). These identity checks are also known as the principles of COHERENCE and SALIENCE, the violations of which are denoted with an asterisk.

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

There exist several variations, however, the most important of which comes from Grosz, Joshi, and Weinstein (1995) who define only one SHIFT transition.[3]

Centering makes two major claims about textual coherence, the first of which is known as Rule 2. Rule 2 states that CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT. Although the Rule was introduced within an algorithm for anaphora resolution, Brennan, Friedman [Walker], and Pollard (1987, p.160) consider it to be relevant to text generation too. Grosz, Joshi, and Weinstein (1995, p.215) also take Rule 2 to suggest that text generation systems should attempt to avoid unfavourable transitions such as SHIFTs.

The second claim, which is implied by the definition of the CB (Poesio et al. 2004), is that CF(U_n) should contain at least one member of CF(U_{n-1}). This became known as the principle of CONTINUITY (Karamanis and Manurung 2002). While Grosz, Joshi, and Weinstein (1995) and Brennan, Friedman [Walker], and Pollard (1987) do not discuss the effect of violating CONTINUITY, Kibble and Power (2000, Figure 1) define the additional transition NOCB to account for this case. Different types of NOCB transitions are introduced by Passoneau (1998) 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).

Kibble (2001) and Beaver (2004) introduced the principles of COHERENCE and SALIENCE, which correspond to the identity checks used to define the transitions (see Table 1). To improve the way Centering resolves pronominal anaphora, Strube and Hahn (1999) introduced a fourth principle called CHEAPNESS and defined as CB(U_n)=CP(U_{n-1}). They also redefined Rule 2 to favour transition pairs which satisfy CHEAPNESS over those which violate it. This means that CHEAPNESS is given priority over every other Centering principle in Strube and Hahn's model.

In addition to the variability caused by the numerous definitions of transitions and the introduction of the various principles, parameters such as "utterance", "ranking" and "realisation" can also be specified in several ways giving rise to different instantiations of Centering (Poesio et al. 2004). The following section discusses how these parameters were defined in the corpora we deploy.

3 "CB(U_{n-1}) undef." in Table 1 stands for the cases where U_{n-1} does not have a CB. Instead of classifying the transition of U_n as a CONTINUE or a RETAIN in such cases, the additional transition ESTABLISHMENT is sometimes used (Kameyama 1998; Poesio et al. 2004).

Table 2
The CF list, the CB, NOCB or Centering transition (see Table 1) and violations of CHEAPNESS (denoted with an asterisk) for each fact in example (1) from the MPIRO-CF corpus.

Fact	CF list: {CP,	next referent}	CB	Transition	CHEAPNESS CB_n=CP_{n-1}
(1a)	{ex1,	amphora}	n.a.	n.a.	n.a.
(1b)	{ex1,	paint-of-kleofr}	ex1	CONTINUE	√
(1c)	{paint-of-kleofr,	en404}	paint-of-kleofr	SMOOTH-SHIFT	*
(1d)	{ex1,	en914}	–	NOCB	n.a.
(1e)	{ex1,	wagner-mus}	ex1	CONTINUE	√
(1f)	{wagner-mus,	germany}	wagner-mus	SMOOTH-SHIFT	*

4. Experimental Data

We made use of the data of Dimitromanolaki and Androutsopoulos (2003), the GNOME corpus (Poesio et al. 2004), and the two corpora that Barzilay and Lapata (2005) experimented with. In this section, we discuss how the Centering representations we utilise were derived from each corpus.

4.1 The MPIRO-CF corpus

Dimitromanolaki and Androutsopoulos (2003, henceforth D&A) derived facts from the database of the MPIRO concept-to-text generation system (Isard et al. 2003), realised them as sentences and organised them in sets. Each set consisted of six facts which were ordered by a domain expert. The orderings produced by this expert were shown to be very close to those produced by two other archeologists (Karamanis and Mellish 2005b).

Our first corpus, MPIRO-CF, consists of 122 orderings that were made available to us by D&A. We computed a CF list for each fact in each ordering by applying the instantiation of Centering introduced by Kibble and Power (2000, 2004) for concept-to-text generation. That is, we took each database fact to correspond to an "utterance" and specified the "realisation" parameter using the arguments of each fact as the members of the corresponding CF list. Table 2 shows the CF lists, the CBs, the Centering transitions and the violations of CHEAPNESS for the following example from MPIRO-CF:

(1) (a) This exhibit is an amphora. (b) This exhibit was decorated by the Painter of Kleofrades. (c) The Painter of Kleofrades used to decorate big vases. (d) This exhibit depicts a warrior performing splachnoscopy before leaving for the battle. (e) This exhibit is currently displayed in the Martin von Wagner Museum. (f) The Martin von Wagner Museum is in Germany.

MPIRO facts consist of two arguments, the first of which was specified as the CP following the definition of "CF ranking" in O'Donnell et al. (2001).[4] Notice that the second argument can often be an entity such as `en914` that is realised by a canned phrase of significant syntactic complexity ("a warrior performing splachnoscopy before leaving for the battle"). Moreover, the deployed definition of "realisation" is similar to what Grosz, Joshi, and Weinstein (1995) call "direct realisation" which ignores potential bridging relations (Clark 1977) between the members of two subsequent CF lists. These

4 This is the main difference between our approach and that of Kibble and Power who allow for more than one potential CP in their CF lists.

Table 3
First two members of the CF list, the CB, NOCB or Centering transition (see Table 1) and violations of CHEAPNESS (denoted with an asterisk) for each finite unit in example (2) from the GNOME-LAB corpus.

Unit	CF list: {CP, next referent}	CB	Transition	CHEAPNESS CB_n=CP_{n-1}
(2a)	{de374, de375}	n.a.	n.a.	n.a.
(2b)	{de376, de374, ... }	de374	RETAIN	$\surd$
(2c)	{de374, de379, ... }	de374	CONTINUE	$*$
(2d)	{de380, de381, ... }	–	NOCB	n.a.

relations are typically not taken into account for information ordering and were not considered in any of the deployed corpora.

4.2 The GNOME-LAB corpus

We also made use of the GNOME corpus (Poesio et al. 2004) which contains object descriptions (museum labels) reliably annotated with features relevant to Centering. The motivation for this study was to examine whether the phenomena observed in MPIRO-CF (which is arguably somewhat artificial) also manifest in texts from the same genre written by humans without the constraints imposed by a text generation system.

Based on the definition of museum labels in Cheng (2002, p.65), we identified 20 such texts in GNOME, which were published in a book and a museum website (and were thus taken to be coherent). The following example is a characteristic text from this subcorpus GNOME-LAB:

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

The GNOME corpus provides us with reliable annotation of discourse units (i.e. clauses and sentences) that can be used for the computation of "utterance" and of NPs which introduce entities to the CF list. Each feature was marked up by at least two annotators and agreement was checked using the κ statistic on part of the corpus.

In order to avoid deviating too much from the MPIRO application domain, we computed the CF lists from the units that seemed to correspond more closely to MPIRO facts. So instead of using *sentence* for the definition of "utterance", we followed most work on Centering for English and computed CF lists from GNOME's *finite units*.[5] The text spans with the indexes (a) to (d) in text (2) are examples of such units. Units such as (2a) are as simple as the MPIRO-generated sentence (1a) while others appear to be of similar syntactic complexity as (1d). On the other hand, the second sentence in (2) consists of two finite units, namely (b) and (c), and appears to correspond to higher degrees of aggregation than is typically seen in an MPIRO fact. The texts in GNOME-LAB consist of 8.35 finite units on average.

5 This definition includes *titles* which do not always have finite verbs, but excludes finite *relative* clauses, the second element of *coordinated VPs* and clause *complements* which are often taken as not having their own CF lists in the Centering literature.

Table 4
Fragment of the entity grid for example (3). The grammatical function of the referents in each sentence is reported using S, O and X (for subject, object and other). The symbol "−" is used for referents which do not occur in the sentence.

	Referents							
Sentences	department	trial	microsoft	evidence	...	products	brands	...
(3a)	S	O	S	X	...	−	−	...
(3b)	−	−	O	−	...	S	O	...
(3c)	−	−	S	O	...	−	−	...
(3d)	−	−	S	−	...	−	−	...
(3e)	−	−	−	−	...	−	−	...
(3f)	−	X	S	−	...	−	−	...

Table 3 shows the first two members of the CF list, the CB, the transitions and the violations of CHEAPNESS for example (2). Note that the same entity (i.e. de374) is used to denote the referent of the NP "Item 144" in (2a) and "its" in (2b), which is annotated as coreferring with "Item 144". All annotated NPs introduce referents to the CF list (which often contains more entities than in MPIRO), but only "direct realisation" is used for the computation of the list. This means that, similarly to the MPIRO domain, bridging relations between e.g. "it" in (2c) and "the terminals" in (2d) are not taken into account.

The members of the CF list were ranked by combining grammatical function with linear order, which is a robust way of estimating "CF ranking" in English (Poesio et al. 2004). In this instantiation, the CP corresponds to the referent of the first NP within the unit that is annotated as a subject or as the post-copular NP in a "there-clause".

4.3 The NEWS and ACCS corpora

Barzilay and Lapata (2005) presented a probabilistic approach for information ordering which is particularly suitable for text-to-text generation and is based on a new representation called the entity grid. A collection of 200 articles from the North American News Corpus (NEWS) and 200 narratives of accidents from the National Transportation Safety Board database (ACCS) was used for training and evaluation. Example (3) presents a characteristic text from the NEWS corpus:

(3) (a) [The Justice Department]$_S$ is conducting [an anti-trust trial]$_O$ against [Microsoft Corp.]$_X$ with [evidence]$_X$ that [the company]$_S$ is increasingly attempting to crush [competitors]$_O$. (b) [Microsoft]$_O$ is accused of trying to forcefully buy into [markets]$_X$ where [its own products]$_S$ are not competitive enough to unseat [established brands]$_O$. (c) [The case]$_S$ revolves around [evidence]$_O$ of [Microsoft]$_S$ aggressively pressuring [Netscape]$_O$ into merging [browser software]$_O$. (d) [Microsoft]$_S$ claims [its tactics]$_S$ are commonplace and good economically. (e) [The government]$_S$ may file [a civil suit]$_O$ ruling that [conspiracy]$_S$ to curb [competition]$_O$ through [collusion]$_X$ is [a violation]$_O$ of [the Sherman Act]$_X$. (f) [Microsoft]$_S$ continues to show [increased earnings]$_O$ despite [the trial]$_X$.

Barzilay and Lapata automatically annotated their corpora for the grammatical function of the NPs in each sentence (denoted in the example by the subscripts S, O and X for subject, object and other, respectively) as well as their coreferential relations (which do not include bridging references). More specifically, they used a parser (Collins 1997) to determine the constituent structure of the sentences from which the grammatical

Table 5
First two members of the CF list, the CB, NOCB or Centering transitions (see Table 1) and violations of CHEAPNESS (denoted with an asterisk) for example (3) from the NEWS corpus.

Sentence	CF list: {CP,	next referents}	CB	Transition	CHEAPNESS CB_n=CP_{n-1}
(3a)	{department,	microsoft, ...}	n.a.	n.a.	n.a.
(3b)	{products,	microsoft, ...}	microsoft	RETAIN	*
(3c)	{microsoft,	case, ...}	microsoft	CONTINUE	*
(3d)	{microsoft,	tactics}	microsoft	CONTINUE	$\surd$
(3e)	{government,	conspiracy, ...}	–	NOCB	n.a.
(3f)	{microsoft,	earnings, ... }	–	NOCB	n.a.

function for each NP was derived.[6] Coreferential NPs such as "Microsoft Corp." and "the company" in (3a) were identified using the system of Ng and Cardie (2002).

The entity grid is a two-dimensional array that captures the distribution of NP referents across sentences in the text using the aforementioned symbols for their grammatical role and the symbol "–" for a referent that does not occur in a sentence. Table 4 illustrates a fragment of the grid for the sentences in example (3).[7]

Barzilay and Lapata use the grid to compute models of coherence which are considerably more elaborate than Centering. To derive an appropriate instantiation of Centering for our investigation, we compute a CF list for each grid row using the referents with the symbols S, O and X. These referents are ranked according to their grammatical function and their position in the text. This definition of "CF ranking" is similar to the one we use in GNOME-LAB. For instance, `department` is ranked higher than `microsoft` in CF(3a) because "the Justice Department" is mentioned before "Microsoft Corp." in the text. The derived sequence of CF lists is used to compute the additional Centering data structures shown in Table 5.

The average number of sentences per text is 10.4 in NEWS and 11.5 in ACCS. As we explain in the next section, our Centering-based metrics of coherence can be deployed directly on unseen texts so we treated all texts in NEWS and ACCS as test data.[8]

5. Computing Centering-based metrics of coherence

Following our previous work (Karamanis and Manurung 2002; Althaus, Karamanis, and Koller 2004), the input to information ordering is an unordered set of information-bearing items represented as CF lists. A set of candidate orderings is produced by creating different permutations of these lists. A metric of coherence uses features from Centering to compute a score for each candidate ordering and select the highest scoring ordering as the output.[9]

6 They also used a small set of patterns to recognise passive verbs and annotate arguments involved in passive constructions with their underlying grammatical function. This is why "Microsoft" is marked with the role O in sentence (3b).

7 If a referent such as `microsoft` is attested by several NPs in the same sentence, e.g. "Microsoft Corp." and "the company" in (3a), the role with the highest priority (in this case S) is used to represent it.

8 By contrast, Barzilay and Lapata used 100 texts in each domain to train their models and reserved the other 100 for testing them.

9 If the best coherence score is assigned to several candidate orderings, then the information ordering algorithm will choose randomly between them.

Table 6
Violations of CONTINUITY (NOCB), COHERENCE, SALIENCE and CHEAPNESS and Centering transitions for example (3), based on the analysis in Table 5. The Table reports the sentences marked with each Centering feature: i.e. sentences (3e) and (3f) are classified as NOCBs, etc.

CONTINUITY∗ NOCB: (3e), (3f)	COHERENCE∗ $CB_n \neq CB_{n-1}$: –	SALIENCE∗ $CB_n \neq CP_n$: (3b)	CHEAPNESS∗ $CB_n \neq CP_{n-1}$: (3b), (3c)
CONTINUE: (3c), (3d)	RETAIN: (3b)	SMOOTH-SHIFT: –	ROUGH-SHIFT: –

A wide range of metrics of coherence can be defined in Centering's terms, simply on the basis of the work we reviewed in section 3. To exemplify this, let us first assume that the ordering in example (3), which is analysed to a sequence of CF lists in Table 5, is a candidate ordering. Table 6 summarises the NOCBs, the violations of COHERENCE, SALIENCE and CHEAPNESS and the Centering transitions for this ordering.[10]

The candidate ordering contains two NOCBs in sentences (3e) and (3f). Its score according to M.NOCB, the metric used by Karamanis and Manurung (2002) and Althaus, Karamanis, and Koller (2004), is 2. Another ordering with fewer NOCBs (should such an ordering exist) will be preferred over this candidate as the selected output of information ordering if M.NOCB is used to guide this process. M.NOCB relies only on CONTINUITY. Since satisfying this principle is a prerequisite for the computation of every other Centering feature, M.NOCB is the simplest possible Centering-based metric and will be used as the baseline in our experiments.

According to Strube and Hahn (1999) the principle of CHEAPNESS is the most important Centering feature for anaphora resolution. We are interested in assessing how suitable M.CHEAP, a metric which utilises CHEAPNESS, is for information ordering. CHEAPNESS is violated twice according to Table 6 so the score of the candidate ordering according to M.CHEAP is 2.[11] If another candidate ordering with fewer violations of CHEAPNESS exists, it will be chosen as a preferred output according to M.CHEAP.

M.BFP employs the transition preferences of Rule 2 as specified by Brennan, Friedman [Walker], and Pollard (1987). The first score to be computed by M.BFP is the sum of CONTINUE transitions, which is 2 for the candidate ordering according to Table 6. If this ordering is found to score higher than every other candidate ordering for the number of CONTINUEs, it is selected as the output. If another ordering is found to have the same number of CONTINUEs, the sum of RETAINs is examined, and so forth for the other two types of Centering transitions.[12]

M.KP, the metric deployed by Kibble and Power (2000) in their text generation system, sums up the NOCBs as well as the violations of CHEAPNESS, COHERENCE and SALIENCE, preferring the ordering with the lowest total cost. In addition to the violations of CONTINUITY and CHEAPNESS, the candidate ordering also violates SALIENCE once so its score according to M.KP is 5. An alternative ordering with a lower score (if any) will be preferred by this metric. Although Kibble and Power (2004)

10 Principles and transitions will be collectively referred to as "features" from now on.
11 In order to estimate the effect of CHEAPNESS only, NOCBs are not counted as violations of CHEAPNESS.
12 Following Brennan, Friedman [Walker], and Pollard (1987), NOCBs are not taken into account for the definition of transitions in M.BFP.

introduced a weighted version of M.KP, the exact weighting of Centering's principles remains an open question as argued by Kibble (2001). This is why we decided to experiment with M.KP instead of its weighted variant.

In the remainder of the paper, we take forward the four metrics motivated in this section as the most appropriate starting point of experimentation. We would like to emphasise, however, that these are not the only possible options. Indeed, similarly to the various ways in which Centering's parameters can be specified, there exist many other ways of using Centering to define metrics of entity coherence for information ordering. These possibilities arise from the numerous other definitions of Centering's transitions and the various ways in which transitions and principles can be combined. These are explored in more detail in Karamanis (2003, Chapter 3), which also provides a formal definition of the metrics discussed above.

6. Evaluation Methodology

Since using naturally occurring discourse in psycholinguistic studies to investigate coherence effects is almost infeasible, computational corpus-based experiments are often the most viable alternative (Poesio et al. 2004; Barzilay and Lee 2004). Corpus-based evaluation can be usefully employed during system development and may be later supplemented by less extended evaluation based on human judgements as suggested in Lapata (2006).

The corpus-based methodology of Karamanis (2003) served as our experimental framework. This methodology is based on the premise that the original sentence order (OSO, Barzilay and Lee 2004) observed in a corpus text is more coherent than any other ordering. If a metric takes an alternative ordering to be more coherent than the OSO, it has to be penalised.

Karamanis (2003) introduced a performance measure called the *classification error rate* which is computed according to the formula: Better(M,OSO)+Equal(M,OSO)/2. Better(M,OSO) stands for the percentage of orderings that score better than the OSO according to a metric M, whilst Equal(M,OSO) is the percentage of orderings that score equal to the OSO.[13] This measure provides an indication of how likely a metric is to lead to an ordering different from the OSO. When comparing several metrics with each other, the one with the lowest classification error rate is the most appropriate for ordering the sentences that the OSO consists of. In other words, *the smaller* the classification error rate, *the better* a metric is expected to perform for information ordering. The *average classification error rate* is used to summarise the performance of each metric in a corpus.

To compute the classification error rate we permute the CF lists of the OSO and classify each alternative ordering as scoring better, equal, or worse than the OSO according to M. When the number of CF lists in the OSO is fairly small, it is feasible to search through all possible orderings. For OSOs consisting of more than 10 CF lists, the classification error rate for the entire population of orderings can be reliably estimated using a random sample of one million permutations (Karamanis 2003, Chapter 5).

13 Weighting Equal(M,OSO) by 0.5 is based on the assumption that, similarly to tossing a coin, the OSO will on average do better than half of the orderings that score the same as it does when other coherence constraints are considered.

Table 7
Average classification error rate for the Centering-based metrics in each corpus.

	Corpus				
Metric	MPIRO-CF	GNOME-LAB	NEWS	ACCS	Mean
M.NOCB	20.42	**19.95**	**30.90**	**15.51**	21.70
M.BFP	**19.91**	33.01	37.90	21.20	28.01
M.KP	53.15	58.22	57.70	55.60	56.12
M.CHEAP	81.04	57.23	64.60	76.29	69.79
N of texts	122	20	200	200	

7. Results

Table 7 shows the average performance of each metric in the corpora employed in our experiments. The smallest, i.e. best, score in each corpus is printed in bold font. The Table indicates that the baseline M.NOCB performs best in three out of four corpora.

The experimental results of the pairwise comparisons of M.NOCB with each of M.CHEAP, M.KP and M.BFP in each corpus are reported in Table 8. The exact number of texts for which the classification error rate of M.NOCB is lower than its competitor for each comparison is reported in the columns headed by "lower". For instance, M.NOCB has a lower classification error rate than M.CHEAP for 110 (out of 122) texts from MPIRO-CF. M.CHEAP achieves a lower classification error rate for just 12 texts, while there do not exist any 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 texts in each corpus, rounded to the third decimal place, is reported in the columns headed by "p".[14] With respect to the exemplified comparison of M.NOCB against M.CHEAP in MPIRO-CF, the p value is lower than 0.001 after rounding (signalled by <0.001 in the corresponding cell). This in turn means that M.NOCB returns a better classification error rate for significantly more texts in MPIRO-CF than M.CHEAP. In other words, M.NOCB outperforms M.CHEAP significantly in this corpus.

Notably, M.NOCB performs significantly better than its competitor in 10 out of 12 cases.[15] In the remaining two comparisons, the difference in performance between M.NOCB and M.BFP is not significant ($p>0.05$). However, this does not constitute evidence against M.NOCB, the simplest of the investigated metrics. In fact, since M.BFP fails to outperform the baseline, the latter may be considered as the most promising solution for information ordering in these cases too by applying Occam's razor. Thus, M.NOCB is shown to be the best performing metric across all four corpora.

8. Discussion

Our experiments show that M.NOCB is the most suitable metric for information ordering among the metrics we experimented with. Despite the differences between our

14 The Sign Test was chosen over its parametric alternatives to test significance because it does not carry specific assumptions about population distributions and variance.
15 This result is significant too according to the two-tailed Sign Test ($p<0.05$).

Table 8
Comparing M.NOCB with M.CHEAP, M.KP and M.BFP in each corpus.

	MPIRO-CF				GNOME-LAB			
	M.NOCB				M.NOCB			
	lower	greater	ties	p	lower	greater	ties	p
M.CHEAP	110	12	0	<0.001	18	2	0	<0.001
M.KP	103	16	3	<0.001	16	2	2	0.002
M.BFP	42	31	49	0.242	12	3	5	0.036
N of texts	122				20			

	NEWS				ACCS			
	M.NOCB				M.NOCB			
	lower	greater	ties	p	lower	greater	ties	p
M.CHEAP	155	44	1	<0.001	183	17	0	<0.001
M.KP	131	68	1	<0.001	167	33	0	<0.001
M.BFP	121	71	8	<0.001	100	100	0	1.000
N of texts	200				200			

corpora (in genre, average length, syntactic complexity, number of referents in the CF list, etc), M.NOCB proves robust across all four of them. It is also the most appropriate metric to use in both application areas we relate our corpora to, namely concept-to-text (MPIRO-CF and GNOME-LAB) as well as text-to-text (NEWS and ACCS) generation. These results indicate that when purely Centering-based metrics are used, simply avoiding NOCBs is more relevant to information ordering than the combinations of additional Centering features that the other metrics make use of.

In this section, we compare our work with other recent evaluation studies, including the corpus-based investigation of Centering by Poesio et al. (2004), discuss the implications of our findings for text generation and summarise our contributions.

8.1 Recent evaluation studies in information ordering

There has been significant recent work on the corpus-based evaluation for information ordering. In this section, we discuss the methodological differences between our work and the studies which are most closely related to it.

Barzilay and Lee (2004) introduce a stochastic model for information ordering which computes the probability of generating the OSO and every alternative ordering. Then, all orderings are ranked according to this probability and the rank given to the OSO is retrieved. Several evaluation measures are discussed, the most of important of which is the *average OSO rank*, i.e. the average rank of the OSOs in their corpora. This measure does not take into account that the OSOs differ in length. However, this information is necessary to estimate reliably the performance of an information ordering approach, as we discuss in Karamanis and Mellish (2005a) in more detail.

Barzilay and Lapata (2005) overcome this difficulty by introducing a performance measure called *ranking accuracy* which expresses the percentage of alternative orderings

that are ranked lower than the OSO. In Karamanis' (2003) terms, ranking accuracy equals $100\% - Better(M, OSO)$, assuming that no equally ranking orderings exist.[16]

Barzilay and Lapata (2005) compare the OSO with just 20 alternative orderings, often sampled out of several millions. On the other hand, Barzilay and Lee (2004) enumerate exhaustively each possible ordering, which might become impractical as the search space grows factorially. We overcame these problems by using a large random sample for the texts which consist of more than 10 sentences as suggested in Karamanis (2003, Chapter 5). Equally important is the emphasis we placed on the use of statistical tests, which were not deployed by either Barzilay and Lee or Barzilay and Lapata.

Lapata (2003) presented a methodology for automatically evaluating generated orderings on the basis of their distance from observed sentence orderings in a corpus. A measure of rank correlation (called Kendall's τ), which was subsequently shown to correlate reliably with human ratings and reading times (Lapata 2006), was used to estimate the distance between orderings.

While τ estimates *how close* the predictions of a metric are to several original orderings, we measure *how likely* a metric is to lead to an ordering different than the OSO. Taking into account more than one OSO for information ordering is the main strength of Lapata's method but to do this one needs to ask several humans to order the same set of sentences (Madnani et al. 2007). Karamanis and Mellish (2005b) conducted an experiment in the MPIRO domain using Lapata's methodology which supplements the work reported in this paper. However, such an approach is less practical for much larger collections of texts such NEWS and ACCS. This is presumably the reason why Barzilay and Lapata (2005) use ranking accuracy instead of τ in their evaluation.

8.2 Previous corpus-based evaluations of Centering

Our work investigates how the coherence score of the OSO compares to the scores of alternative orderings of the sentences that the OSO consists of. As Kibble (2001, p.582) noticed, this question is crucial from an information ordering viewpoint, but was not taken into account by any previous corpus-based study of Centering. Grosz, Joshi, and Weinstein (1995, p.215) also suggested that Rule 2 should be tested by examining "alternative multi-utterance sequences that differentially realize the same content". We are first to have pursued this research objective in the evaluation of Centering for information ordering.

Poesio et al. (2004) observed that there remained a large number of NOCBs under every instantiation of Centering they tested and concluded that Centering is inadequate as a coherence model.[17] However, the frequency of NOCBs does not necessarily provide adequate indication of how appropriate NOCBs (and Centering in general) are for information ordering. Although over 50% of the transitions in GNOME-LAB are NOCBs, the average classification error rate of approximately 20% for M.NOCB suggests that the OSO tends to be in greater agreement with the preference to avoid NOCBs than 80% of the alternative orderings. Thus, it appears that the observed ordering in the corpus does optimise with respect to the number of potential NOCBs to a great extent.

16 Neither Barzilay and Lapata (2005) nor Barzilay and Lee (2004) appear to consider the possibility that two orderings may be equally ranked.

17 We viewed the definition of the Centering instantiation as being related to the application domain, as we explained in section 4. This is why, unlike Poesio et al., we did not experiment with different instantiations of Centering on the same data.

8.3 A simple and robust baseline for text generation

How likely is M.NOCB to come up with the attested ordering in the corpus (the OSO) if it is actually used to guide an algorithm that orders the CF lists in our corpora? The average classification error rates (Table 7) estimate exactly this variable. The performance of M.NOCB varies across the corpora from about 15.5% (ACCS) to 30.9% (NEWS). We attribute this variation to the aforesaid differences between the corpora. Notice, however, that these differences affect all metrics in a similar way, not allowing for another metric to significantly outperform M.NOCB.

Noticeably, even in ACCS, for which M.NOCB achieves its best performance, approximately one out of six alternative orderings on average are taken to be more coherent than the OSO. Given the average number of sentences per text in this corpus (11.5), this means that several millions of alternative orderings are often taken to be more coherent than the gold standard.

Barzilay and Lapata (2005) report an average ranking accuracy of 87.3% for their best sentence ordering method in ACCS. This corresponds to an average classification error rate of 12.7% (assuming that there are no equally scoring orderings in their evaluation: see section 8.1). This is equal to an improvement of just 2.8% over the performance of our baseline metric (15.5%) using a coherence model which is substantially more elaborate than Centering. However, it is in NEWS (for which M.NOCB returns its worse performance of 30.9%) that this model shows its real strength approximating an average classification error rate of 9.6%, which corresponds to an improvement of 21.3% over our baseline. We believe that the experiments reported in this paper put the studies of our colleagues in better perspective by providing a reliable baseline to compare their metrics against.

8.4 Moving beyond Centering-based metrics

Following McKeown (1985), Kibble and Power argue in favour of an integrated approach for concept-to-text generation in which the same Centering features are used at different stages in the generation pipeline. However, our study suggests that features such as CHEAPNESS and the Centering transitions are not particularly relevant to information ordering. The poor performance of these features can be explained by the fact that they were originally introduced to account for pronoun resolution rather than information ordering. CONTINUITY, on the other hand, captures a fundamental intuition about entity coherence which constitutes part of several other discourse theories.[18]

Yet, CONTINUITY captures just one aspect of coherence. This explains the relatively high classification error rates for M.NOCB, which needs to be supplemented with other coherence-inducing factors in order to be used in practice. This verifies the premises of researchers such as Kibble and Power who a priori use features derived from Centering in combination with other factors in the definition of their metrics. Our work should be quite helpful for that effort too, suggesting that M.NOCB is a better starting point for defining such metrics than M.CHEAP or M.KP.

18 We thank one anonymous reviewer for suggesting this explanation of our results.

9. Conclusion

In conclusion, our analysis sheds more light on two previously unaddressed questions in the corpus-based evaluation of Centering: (i) which aspects of Centering are most relevant to information ordering and (ii) to what extent Centering on its own can be useful for this purpose. We have shown that the metric which relies exclusively on NOCB transitions (M.NOCB) sets a baseline that cannot be outperformed by other coherence metrics which make use of additional Centering features. Although this metric does not perform well enough to be used on its own, it constitutes a simple, yet robust, baseline against which more elaborate information ordering approaches can be tested during system development in both text-to-text and concept-to-text generation.

This work can be extended in numerous ways. For instance, given the abundance of possible Centering-based metrics one may investigate whether a different metric can outperform M.NOCB in any corpus or application domain. M.NOCB can also serve as the starting point for the definition of more informed metrics which will incorporate additional coherence-inducing factors. Finally, given that we used the instantiation of Centering which seemed to correspond more closely to the targeted application domains, the extent to which computing the CF list in a different way may affect the performance of the metrics is another question to explore in future work.

Acknowledgments

Many thanks to Aggeliki Dimitromanolaki, Mirella Lapata and Regina Barzilay for their data, to David Schlangen, Ruli Manurung, James Soutter and Le An Ha for programming solutions and to Ruth Seal and two anonymous reviewers for their comments. Nikiforos Karamanis received support by the Greek State Scholarships Foundation (IKY) as a PhD student in Edinburgh as well as the Rapid Item Generation project and the BBSRC-funded FlySlip grant (No 38688) as a postdoc in Wolverhampton and Cambridge respectively.

References

Althaus, Ernst, Nikiforos Karamanis, and Alexander Koller. 2004. Computing locally coherent discourses. In *Proceedings of ACL 2004*, pages 399–406.

Barzilay, Regina, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Barzilay, Regina and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of ACL 2005*, pages 141–148.

Barzilay, Regina and Lillian Lee. 2004. Catching the drift: Probabilistic content models with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, pages 113–120.

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

Bollegala, Danushka, Naoaki Okazaki, and Mitsuru Ishizuka. 2006. A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of ACL-COLING 2006*, pages 385–392.

Brennan, Susan E., Marilyn A. Friedman [Walker], and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of ACL 1987*, pages 155–162.

Cheng, Hua. 2002. *Modelling Aggregation Motivated Interactions in Descriptive Text Generation*. Ph.D. thesis, Division of Informatics, University of Edinburgh.

Clark, Herbert. H. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge University Press, pages 9–27.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL-EACL 1997*, pages 16–23.

Dimitromanolaki, Aggeliki and Ion Androutsopoulos. 2003. Learning to order facts for discourse planning in natural language generation. In *Proceedings of ENLG 2003*, pages 23–30.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework

for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
Hovy, Eduard. 1988. Planning coherent multisentential text. In *Proceedings of ACL 1988*, pages 163–169.
Isard, Amy, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45.
Ji, Paul and Stephen Pulman. 2006. Sentence ordering with manifold-based classification in multi-document summarization. In *Proceedings of EMNLP 2006*, pages 526–533.
Kameyama, Megumi. 1998. Intrasentential centering: A case study. In Walker et al. (Walker, Joshi, and Prince 1998), pages 89–122.
Kan, Min-Yen and Kathleen McKeown. 2002. Corpus-trained text generation for summarization. In *Proceedings of INLG 2002*, pages 1–8.
Karamanis, N. 2006. Evaluating centering for information ordering in two new domains. In *Proceedings of NAACL 2006, Companion Volume*, pages 65–68.
Karamanis, N., M. Poesio, C. Mellish, and J. Oberlander. 2004. Evaluating centering-based metrics of coherence using a reliably annotated corpus. In *Proceedings of ACL 2004*, pages 391–398.
Karamanis, Nikiforos. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, Division of Informatics, University of Edinburgh.
Karamanis, Nikiforos and Hisar Maruli Manurung. 2002. Stochastic text structuring using the principle of continuity. In *Proceedings of INLG 2002*, pages 81–88.
Karamanis, Nikiforos and Chris Mellish. 2005a. 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*, pages 13–18.
Karamanis, Nikiforos and Chris Mellish. 2005b. Using a corpus of sentence orderings defined by many experts to evaluate metrics of coherence for text structuring. In *Proceedings of ENLG 2005*, pages 174–179.
Kibble, Rodger. 2001. A reformulation of rule 2 of centering theory. *Computational Linguistics*, 27(4):579–587.
Kibble, Rodger and Richard Power. 2000. An integrated framework for text planning and pronominalisation. In *Proceedings of INLG 2000*, pages 77–84.
Kibble, Rodger and Richard Power. 2004. Optimizing referential coherence in text generation. *Computational Linguistics*, 30(4):401–416.
Knott, Alistair, Jon Oberlander, Mick O'Donnell, and Chris Mellish. 2001. Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*. John Benjamins, Amsterdam, chapter 7, pages 181–196.
Lapata, Mirella. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, pages 545–552.
Lapata, Mirella. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1–14.
Madnani, Nitin, Rebecca Passonneau, Necip Fazil Ayan, John Conroy, Bonnie Dorr, Judith Klavans, Dianne O'Leary, and Judith Schlesinger. 2007. Measuring variability in sentence ordering for news summarization. In *Proceedings of ENLG 2007*, pages 81–88.
Mann, William C. and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organisation. Technical Report RR-87-190, University of Southern California / Information Sciences Institute.
Marcu, Daniel. 1997. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto.
McKeown, Kathleen. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Studies in Natural Language Processing. Cambridge University Press.
Mellish, Chris, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of INLG 1998*, pages 98–107.
Miltsakaki, Eleni and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.
Ng, Vincent and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL 2002*, pages 104–111.
O'Donnell, Mick, Chris Mellish, Jon Oberlander, and Alistair Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language*

Engineering, 7(3):225–250.
Passoneau, Rebecca J. 1998. Interaction of discourse structure with explicitness of discourse anaphoric phrases. In Walker et al. (Walker, Joshi, and Prince 1998), pages 327–358.
Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: a parametric theory and its instantiations. Technical Report CSM-369, Department of Computer Science, University of Essex. Extended version of the paper that appeared in *Computational Linguistics* 30(3):309-363, 2004.
Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
Scott, Donia and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*. Academic Press, pages 47–74.
Siddharthan, Advaith. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.
Sidner, Candace L. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English*. Ph.D. thesis, AI Laboratory/MIT, Cambridge, MA, June. Also available as Technical Report No. AI-TR-537.
Soricut, Radu and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of ACL-COLING 2006 Poster Session*, pages 803–810.
Strube, Michael and Udo Hahn. 1999. Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.
Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince, editors. 1998. *Centering Theory in Discourse*. Clarendon Press, Oxford.