Evaluating Centering-based metrics of coherence for text structuring using a reliably annotated corpus

Nikiforos Karamanis,[♣] Massimo Poesio,[♢] Chris Mellish,[♠] and Jon Oberlander[♣]
[♣]School of Informatics, University of Edinburgh, UK, {nikiforo,jon}@ed.ac.uk
[♢]Dept. of Computer Science, University of Essex, UK, poesio at essex dot ac dot uk
[♠]Dept. of Computing Science, University of Aberdeen, UK, cmellish@csd.abdn.ac.uk

Abstract

We use a reliably annotated corpus to compare metrics of coherence based on Centering Theory with respect to their potential usefulness for text structuring in natural language generation. Previous corpus-based evaluations of the coherence of text according to Centering did not compare the coherence of the chosen text structure with that of the possible alternatives. A corpus-based methodology is presented which distinguishes between Centering-based metrics taking these alternatives into account, and represents therefore a more appropriate way to evaluate Centering from a text structuring perspective.

1 Motivation

Our research area is descriptive text generation (O'Donnell et al., 2001; Isard et al., 2003), i.e. the generation of descriptions of objects, typically museum artefacts, depicted in a picture. Text (1), from the GNOME corpus (Poesio et al., 2004), is an example of short human-authored text from this genre:

(1) (a) 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.

According to Centering Theory (Grosz et al., 1995; Walker et al., 1998a), an important factor for the felicity of (1) is its entity coherence: the way CENTERS (discourse entities), such as the referent of the NPs "144" in clause (a) and "its" in clause (b), are introduced and discussed in subsequent clauses. It is often claimed in current work on in natural language generation that the constraints on felicitous text proposed by the theory are useful to guide text structuring, in combination with other factors (see (Karamanis, 2003) for an overview). However, how successful Centering's constraints are *on their own* in generating a felicitous text structure is an open question, already raised by the seminal papers of the theory (Brennan et al., 1987; Grosz et al., 1995). In this work, we explored this question by developing an approach to text structuring purely based on Centering, in which the role of other factors is deliberately ignored.

In accordance with recent work in the emerging field of text-to-text generation (Barzilay et al., 2002; Lapata, 2003), we assume that the input to text structuring is a set of clauses. The output of text structuring is merely an *ordering* of these clauses, rather than the tree-like structure of database facts often used in traditional deep generation (Reiter and Dale, 2000). Our approach is further characterized by two key insights. The first distinguishing feature is that we assume a *search-based* approach to text structuring (Mellish et al., 1998; Kibble and Power, 2000; Karamanis and Manurung, 2002) in which many candidate orderings of clauses are evaluated according to scores assigned by a given metric, and the best-scoring ordering among the candidate solutions is chosen. The second novel aspect is that our approach is based on the position that the most straightforward way of using Centering for text structuring is by defining a Centering-based *metric* of coherence Karamanis (2003). Together, these two assumptions lead to a view of text planning in which the constraints of Centering act not as *filters*, but as *ranking factors*, and the text planner may be forced to choose a sub-optimal solution.

However, Karamanis (2003) pointed out that many metrics of coherence can be derived from the claims of Centering, all of which could be used for the type of text structuring assumed in this paper. Hence, a general methodology for identifying which of these metrics represent the most promising candidates for text structuring is required, so that at least some of them can be compared empirically. This is the second research question that this paper addresses, building upon previous work on corpus-based evaluations of Centering, and particularly the methods used by Poesio et al. (2004). We use the GNOME corpus (Poesio et al., 2004) as the domain of our experiments because it is reliably annotated with features relevant to Centering and contains the genre that we are mainly interested in.

To sum up, in this paper we try to identify the most promising Centering-based metric for text structuring, and to evaluate how useful this metric is for that purpose, using corpus-based methods instead of generally more expensive psycholinguistic techniques. The paper is structured as follows. After discussing how the GNOME corpus has been used in previous work to evaluate the coherence of a text according to Centering we discuss why such evaluations are not sufficient for text structuring. We continue by showing how Centering can be used to define different metrics of coherence which might be useful to drive a text planner. We then outline a corpus-based methodology to choose among these metrics, estimating how well they are expected to do when used by a text planner. We conclude by discussing our experiments in which this methodology is applied using a subset of the GNOME corpus.

2 Evaluating the coherence of a corpus text according to Centering

In this section we briefly introduce Centering, as well as the methodology developed in Poesio et al. (2004) to evaluate the coherence of a text according to Centering.

2.1 Computing CF lists, CPs and CBs

According to Grosz et al. (1995), each "utterance" in a discourse is assigned a list of forward looking centers (CF list) each of which is "realised" by at least one NP in the utterance. The members of the CF list are "ranked" in order of prominence, the first element being the preferred center CP.

In this paper, we used what we considered to be the most common definitions of the central notions of Centering (its 'parameters'). Poesio et al. (2004) point out that there are many definitions of parameters such as "utterance", "ranking" or "realisation", and that the setting of these parameters greatly affects the predictions of the theory;[1] however, they found violations of the Centering constraints with any way of setting the parameters (for instance, at least 25% of utterances have no CB under any such setting), so that the questions addressed by our work arise for all other settings as well.

Following most mainstream work on Centering for English, we assume that an "utterance" corresponds to what is annotated as a *finite unit* in the GNOME corpus.[2] The spans of text with the indexes (a) to (d) in example (1) are examples. This definition of utterance is not optimal from the point of view of minimizing Centering violations (Poesio et al., 2004), but in this way most utterances are the realization of a single proposition; i.e., the impact of aggregation is greatly reduced. Similarly, we use grammatical function (`gf`) combined with linear order within the unit (what Poesio et al. (2004) call *gftherelin*) for CF ranking. In this configuration, the CP is the referent of the first NP within the unit that is annotated as a subject for its `gf`.[3]

Example (2) shows the relevant annotation features of unit `u210` which corresponds to utterance (a) in example (1). According to *gftherelin*, the CP of (a) is the referent of `ne410` "144".

(2)
```
<unit finite='finite-yes' id='u210'>

<ne id="ne410" gf="subj">144</ne>

is

<ne id="ne411" gf="predicate">
a torc</ne> </unit>.
```

The ranking of the CFs other than the CP is defined according to the following preference on their `gf` (Brennan et al., 1987): `obj`>`iobj`>`other`. CFs with the same `gf` are ranked according to the linear order of the corresponding NPs in the utterance. The second column of Table 1 shows how the utterances in example (1) are automatically translated by the scripts developed by Poesio et al. (2004) into a

[1] For example, one could equate "utterance" with *sentence* (Strube and Hahn, 1999; Miltsakaki, 2002), use *indirect realisation* for the computation of the CF list (Grosz et al., 1995), rank the CFs according to their *information status* (Strube and Hahn, 1999), etc.

[2] Our definition includes *titles* which are not always finite units, 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 literature.

[3] Or as a post-copular subject in a *there*-clause.

U	CF list: {CP,	 other CFs}	CB	Transition	CHEAPNESS $CB_n=CP_{n-1}$
(a)	{de374,	de375}	n.a.	n.a.	n.a.
(b)	{de376,	de374, de377}	de374	RETAIN	+
(c)	{de374,	de379}	de374	CONTINUE	∗
(d)	{de380,	de381, de382}	-	NOCB	+

Table 1: CP, CFs other than CP, CB, NOCB or standard (see Table 2) transition and violations of CHEAPNESS (denoted with an asterisk) for each utterance (U) in example (1)

	COHERENCE: $CB_n=CB_{n-1}$ or NOCB in CF_{n-1}	COHERENCE∗: $CB_n \neq CB_{n-1}$
SALIENCE: $CB_n=CP_n$	CONTINUE	SMOOTH-SHIFT
SALIENCE∗: $CB_n \neq CP_n$	RETAIN	ROUGH-SHIFT

Table 2: COHERENCE, SALIENCE and the table of standard transitions

sequence of CF lists, each decomposed into the CP and the CFs other than the CP, according to the chosen setting of the Centering parameters. Note that the CP of (a) is the center de374 and that the same center is used as the referent of the other NPs which are annotated as coreferring with ne410.

Given two subsequent utterances U_{n-1} and U_n, with CF lists CF_{n-1} and CF_n respectively, the backward looking center of U_n, CB_n, is defined as the highest ranked element of CF_{n-1} which also appears in CF_n (Centering's Constraint 3). For instance, the CB of (b) is de374. The third column of Table 1 shows the CB for each utterance in (1).[4]

2.2 Computing transitions

As the fourth column of Table 1 shows, each utterance, with the exception of (a), is also marked with a transition from the previous one. When CF_n and CF_{n-1} do not have any centers in common, we compute the NOCB transition (Kibble and Power, 2000) (Poesio *et al*'s NULL transition) for U_n (e.g., utterance (d) in Table 1).[5]

Following again the terminology in Kibble and Power (2000), we call the requirement that CB_n be the same as CB_{n-1} the principle of COHERENCE and the requirement that CB_n be the same as CP_n the principle of SALIENCE. Each of these principles can be satisfied or violated while their various combinations give rise to the standard transitions of Centering shown in Table 2; Poesio *et al*'s scripts compute these violations.[6] We also make note of the preference between these transitions, known as Centering's Rule 2 (Brennan et al., 1987): CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT.

Finally, the scripts determine whether CB_n is the same as CP_{n-1}, known as the principle of CHEAPNESS (Strube and Hahn, 1999). The last column of Table 1 shows the violations of CHEAPNESS (denoted with an asterisk) in (1).[7]

2.3 Evaluating the coherence of a text and text structuring

The statistics about transitions computed as just discussed can be used to determine the degree to which a text conforms with, or violates, Centering's principles. Poesio et al. (2004) found that NOCBs account for more than 50%

[4] In accordance with Centering, no CB is computed for (a), the first utterance in the sequence.

[5] In this study we do not take *indirect realisation* into account, i.e., we ignore the bridging reference (annotated in the corpus) between the referent of "it" de374 in (c) and the referent of "the terminals" de380 in (d), by virtue of which de374 might be thought as being a member of the CF list of (d). Poesio et al. (2004) showed that hypothesizing indirect realization eliminates many violations of ENTITY CONTINUITY, the part of Constraint 1 that rules out NOCB transitions. However, in this work we are treating CF lists as an abstract representation of the atomic facts the algorithm has to structure, i.e., we are assuming that CFs are arguments of such facts; including indirectly realized entities in CF lists would violate this assumption.

[6] If the second utterance in a sequence U_2 has a CB, then it is taken to be either a CONTINUE or a RETAIN, although U_1 is not classified as a NOCB.

[7] As for the other two principles, no violation of CHEAPNESS is computed for (a) or when U_n is marked as a NOCB.

of the transitions in the GNOME corpus in configurations such as the one used in this paper. More generally, a significant percentage of NOCBs (at least 20%) and other "dispreferred" transitions was found with all parameter configurations tested by Poesio et al. (2004) and indeed by all previous corpus-based evaluations of Centering such as Passoneau (1998), Di Eugenio (1998), Strube and Hahn (1999) among others. These results led Poesio et al. (2004) to the conclusion that the entity coherence as formalized in Centering should be supplemented with an account of other coherence inducing factors to explain what makes texts coherent.

These studies, however, do not investigate the question that is most important from the text structuring perspective adopted in this paper: whether there would be alternative ways of structuring the text that would result in fewer violations of Centering's constraints (Kibble, 2001). Consider the NOCB utterance (d) in (1). Simply observing that this transition is 'dispreferred' ignores the fact that every other ordering of utterances (b) to (d) would result in *more* NOCBs than those found in (1). Even a text-structuring algorithm functioning solely on the basis of the Centering constraints might therefore still choose the particular order in (1). In other words, a metric of text coherence purely based on Centering principles–trying to minimize the number of NOCBs–may be sufficient to explain why this order of clauses was chosen, at least in this particular genre, without need to involve more complex explanations. In the rest of the paper, we consider several such metrics, and use the texts in the GNOME corpus to choose among them. We return to the issue of coherence (i.e., whether additional coherence-inducing factors need to be stipulated in addition to those assumed in Centering) in the Discussion.

3 Centering-based metrics of coherence

As said previously, we assume a text structuring system taking as input a set of utterances represented in terms of their CF lists. The system orders these utterances by applying a bias in favour of the best scoring ordering among the candidate solutions for the preferred output.[8] In this section, we discuss how the Centering concepts just described can be used to define metrics of coherence which might be useful for text structuring.

The simplest way to define a metric of coherence using notions from Centering is to classify each ordering of propositions according to the number of NOCBs it contains, and pick the ordering with the fewest NOCBs. We call this metric M.NOCB, following (Karamanis and Manurung, 2002). Because of its simplicity, M.NOCB serves as the baseline metric in our experiments.

We consider three more metrics. M.CHEAP is biased in favour of the ordering with the fewest violations of CHEAPNESS. M.KP sums up the NOCBs and the violations of CHEAPNESS, COHERENCE and SALIENCE, preferring the ordering with the lowest total cost (Kibble and Power, 2000). Finally, M.BFP employs the preferences between standard transitions as expressed by Rule 2. More specifically, M.BFP selects the ordering with the highest number of CONTINUEs. If there exist several orderings which have the most CONTINUEs, the one which has the most RETAINs is favoured. The number of SMOOTH-SHIFTs is used only to distinguish between the orderings that score best for CONTINUEs as well as for RETAINs, etc.

In the next section, we present a general methodology to compare these metrics, using the actual ordering of clauses in real texts of a corpus to identify the metric whose behavior mimics more closely the way these actual orderings were chosen. This methodology was implemented in a program called the *System for Evaluating Entity Coherence* (SEEC).

4 Exploring the space of possible orderings

In section 2, we discussed how an ordering of utterances in a text like (1) can be translated into a sequence of CF lists, which is the representation that the Centering-based metrics operate on. We use the term *Basis for Comparison* (BfC) to indicate this sequence of CF lists. In this section, we discuss how the BfC is used in our search-oriented evaluation methodology to calculate a performance measure for each metric and compare them with each other. In the next section, we will see how our corpus was used to identify the most promising Centering-based metric for a text classifier.

4.1 Computing the classification rate

The performance measure we employ is called the *classification rate* of a metric M on a cer-

[8] Additional assumptions for choosing between the orderings that are assigned the best score are presented in the next section.

tain BfC B. The classification rate estimates the ability of M to produce B as the output of text structuring according to a specific generation scenario.

The first step of SEEC is to search through the space of possible orderings defined by the permutations of the CF lists that B consists of, and to divide the explored search space into sets of orderings that score better, equal, or worse than B according to M.

Then, the classification rate is defined according to the following generation scenario. We assume that an ordering has higher chances of being selected as the output of text structuring the better it scores for M. This is turn means that the fewer the members of the set of better scoring orderings, the better the chances of B to be the chosen output.

Moreover, we assume that additional factors play a role in the selection of one of the orderings that score the same for M. On average, B is expected to sit in the middle of the set of equally scoring orderings with respect to these additional factors. Hence, half of the orderings with the same score will have better chances than B to be selected by M.

The classification rate υ of a metric M on B expresses the expected percentage of orderings with a higher probability of being generated than B according to the scores assigned by M and the additional biases assumed by the generation scenario as follows:

(3) Classification rate:

$$\upsilon(M,B) = Better(M) + \frac{Equal(M)}{2}$$

$Better(M)$ stands for the percentage of orderings that score better than B according to M, whilst $Equal(M)$ is the percentage of orderings that score equal to B according to M. If $\upsilon(M_x,B)$ is the classification rate of M_x on B, and $\upsilon(M_y,B)$ is the classification rate of M_y on B, M_y is a more suitable candidate than M_x for generating B if $\upsilon(M_y,B)$ is smaller than $\upsilon(M_x,B)$.

4.2 Generalising across many BfCs

In order for the experimental results to be reliable and generalisable, M_x and M_y should be compared on more than one BfC from a corpus C. In our standard analysis, the BfCs $B_1,...,B_m$ from C are treated as the random factor in a repeated measures design since each BfC contributes a score for each metric. Then, the classification rates for M_x and M_y on the BfCs are compared with each other and significance is tested using the Sign Test. After calculating the number of BfCs that return a lower classification rate for M_x than for M_y and vice versa, the Sign Test reports whether the difference in the number of BfCs is significant, that is, whether there are significantly more BfCs with a lower classification rate for M_x than the BfCs with a lower classification rate for M_y (or vice versa).[9]

Finally, we summarise the performance of M on m BfCs from C in terms of the *average classification rate* Y:

(4) Average classification rate:

$$Y(M,C) = \frac{\upsilon(M,B_1)+...+\upsilon(M,B_m)}{m}$$

5 Using the GNOME corpus for a search-based comparison of metrics

We will now discuss how the methodology discussed above was used to compare the Centering-based metrics discussed in Section 3, using the original ordering of texts in the GNOME corpus to compute the average classification rate of each metric.

The GNOME corpus contains texts from different genres, not all of which are of interest to us. In order to restrict the scope of the experiment to the text-type most relevant to our study, we selected 20 "museum labels", i.e., short texts that describe a concrete artefact, which served as the input to SEEC together with the metrics in section 3.[10]

5.1 Permutation and search strategy

In specifying the performance of the metrics we made use of a simple permutation heuristic exploiting a piece of domain-specific communication knowledge (Kittredge et al., 1991). Like Dimitromanolaki and Androutsopoulos (2003), we noticed that utterances like (a) in example (1), should always appear at the beginning of a felicitous museum label. Hence, we restricted the orderings considered by the SEEC

[9] The Sign Test was chosen over its parametric alternatives to test significance because it does not carry specific assumptions about population distributions and variance. It is also more appropriate for small samples like the one used in this study.

[10] Note that example (1) is characteristic of the genre, not the length, of the texts in our subcorpus. The number of CF lists that the BfCs consist of ranges from 4 to 16 (average cardinality: 8.35 CF lists).

Pair	M.NOCB			p	Winner
	lower	greater	ties		
M.NOCB vs M.CHEAP	18	2	0	0.000	M.NOCB
M.NOCB vs M.KP	16	2	2	0.001	M.NOCB
M.NOCB vs M.BFP	12	3	5	0.018	M.NOCB
N	20				

Table 3: Comparing M.NOCB with M.CHEAP, M.KP and M.BFP in GNOME

to those in which the first CF list of B, CF_1, appears in first position.[11]

For very short texts like (1), which give rise to a small BfC, the search space of possible orderings can be enumerated exhaustively. However, when B consists of many more CF lists, it is impractical to explore the search space in this way. Elsewhere we show that even in these cases it is possible to estimate $v(M, B)$ reliably for the whole population of orderings using a large random sample. In the experiments reported here, we had to resort to random sampling only once, for a BfC with 16 CF lists.

5.2 Comparing M.NOCB with other metrics

The experimental results of the comparisons of the metrics from section 3, computed using the methodology in section 4, are reported in Table 3.

In this table, the baseline metric M.NOCB is compared with each of M.CHEAP, M.KP and M.BFP. The first column of the Table identifies the comparison in question, e.g. M.NOCB versus M.CHEAP. The exact number of BfCs for which the classification rate of M.NOCB is lower than its competitor for each comparison is reported in the next column of the Table. For example, M.NOCB has a lower classification rate than M.CHEAP for 18 (out of 20) BfCs from the GNOME corpus. M.CHEAP only achieves a lower classification rate for 2 BfCs, and there are no ties, i.e. cases where the classification rate of the two metrics is the same. The p value returned by the Sign Test for the difference in the number of BfCs, rounded to the third decimal place, is reported in the fifth column of the Table. The last column of the Table 3 shows M.NOCB as the "winner" of the comparison with M.CHEAP since it has a lower classification rate than its competitor for significantly more BfCs in the corpus.[12]

Overall, the Table shows that M.NOCB does significantly better than the other three metrics which employ additional Centering concepts. This result means that there exist proportionally fewer orderings with a higher probability of being selected than the BfC when M.NOCB is used to guide the hypothetical text structuring algorithm instead of the other metrics.

Hence, M.NOCB is the most suitable among the investigated metrics for structuring the CF lists in GNOME. This in turn indicates that simply avoiding NOCB transitions is more relevant to text structuring than the combinations of the other Centering notions that the more complicated metrics make use of. (However, these notions might still be appropriate for other tasks, such as anaphora resolution.)

6 Discussion: the performance of M.NOCB

We already saw that Poesio et al. (2004) found that the majority of the recorded transitions in the configuration of Centering used in this study are NOCBs. However, we also explained in section 2.3 that what really matters when trying to determine whether a text might have been generated only paying attention to Centering constraints is the extent to which it would be possible to 'improve' upon the ordering chosen in that text, given the information that the text structuring algorithm had to convey. The average classification rate of M.NOCB is an esti-

[11] Thus, we assume that when the set of CF lists serves as the input to text structuring, CF_1 will be identified as the initial CF list of the ordering to be generated using annotation features such as the *unit type* which distinguishes (a) from the other utterances in (1).

[12] No winner is reported for a comparison when the p value returned by the Sign Test is not significant (ns), i.e. greater than 0.05. Note also that despite conducting more than one pairwise comparison simultaneously we refrain from further adjusting the overall threshold of significance (e.g. according to the Bonferroni method, typically used for multiple planned comparisons that employ parametric statistics) since it is assumed that choosing a conservative statistic such as the Sign Test already provides substantial protection against the possibility of a type I error.

Pair	M.NOCB			p	Winner
	lower	greater	ties		
M.NOCB vs M.CHEAP	110	12	0	0.000	M.NOCB
M.NOCB vs M.KP	103	16	3	0.000	M.NOCB
M.NOCB vs M.BFP	41	31	49	0.121	ns
N	122				

Table 4: Comparing M.NOCB with M.CHEAP, M.KP and M.BFP using the novel methodology in MPIRO

mate of exactly this variable, indicating whether M.NOCB is likely to arrive at the BfC during text structuring.

The average classification rate Y for M.NOCB on the subcorpus of GNOME studied here, for the parameter configuration of Centering we have assumed, is 19.95%. This means that on average the BfC is close to the top 20% of alternative orderings when these orderings are ranked according to their probability of being selected as the output of the algorithm.

On the one hand, this result shows that although the ordering of CF lists in the BfC might not completely minimise the number of observed NOCB transitions, the BfC tends to be in greater agreement with the preference to avoid NOCBs than most of the alternative orderings. In this sense, it appears that the BfC optimises with respect to the number of potential NOCBs to a certain extent. On the other hand, this result indicates that there are quite a few orderings which would appear more likely to be selected than the BfC.

We believe this finding can be interpreted in two ways. One possibility is that M.NOCB needs to be supplemented by other features in order to explain why the original text was structured this way. This is the conclusion arrived at by Poesio et al. (2004) and those text structuring practitioners who use notions derived from Centering in combination with other coherence constraints in the definitions of their metrics. There is also a second possibility, however: we might want to reconsider the assumption that human text planners are trying to ensure that each utterance in a text is locally coherent. They might do all of their planning just on the basis of Centering constraints, at least in this genre –perhaps because of resource limitations– and simply accept a certain degree of incoherence. Further research on this issue will require psycholinguistic methods; our analysis nevertheless sheds more light on two previously unaddressed questions in the corpus-based evaluation of Centering – a) which of the Centering notions are most relevant to the text structuring task, and b) to which extent Centering on its own can be useful for this purpose.

7 Further results

In related work, we applied the methodology discussed here to a larger set of existing data (122 BfCs) derived from the MPIRO system and ordered by a domain expert (Dimitromanolaki and Androutsopoulos, 2003). As Table 4 shows, the results from MPIRO verify the ones reported here, especially with respect to M.KP and M.CHEAP which are overwhelmingly beaten by the baseline in the new domain as well. Also note that since M.BFP fails to overtake M.NOCB in MPIRO, the baseline can be considered the most promising solution among the ones investigated in both domains by applying Occam's logical principle.

We also tried to account for some additional constraints on coherence, namely local rhetorical relations, based on some of the assumptions in Knott et al. (2001), and what Karamanis (2003) calls the "PageFocus" which corresponds to the main entity described in a text, in our example `de374`. These results, reported in (Karamanis, 2003), indicate that these constraints conflict with Centering as formulated in this paper, by increasing - instead of reducing - the classification rate of the metrics. Hence, it remains unclear to us how to improve upon M.NOCB.

In our future work, we would like to experiment with more metrics. Moreover, although we consider the parameter configuration of Centering used here a plausible choice, we intend to apply our methodology to study different instantiations of the Centering parameters, e.g. by investigating whether "indirect realisation" reduces the classification rate for M.NOCB compared to "direct realisation", etc.

Acknowledgements

Special thanks to James Soutter for writing the program which translates the output produced by GNOME's scripts into a format appropriate for SEEC. The first author was able to engage in this research thanks to a scholarship from the Greek State Scholarships Foundation (IKY).

References

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

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

Barbara Di Eugenio. 1998. Centering in Italian. In Walker et al. (Walker et al., 1998b), pages 115–137.

Aggeliki Dimitromanolaki and Ion Androutsopoulos. 2003. Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest, Hungary.

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

Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Colin Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45.

Nikiforos Karamanis and Hisar Maruli Manurung. 2002. Stochastic text structuring using the principle of continuity. In *Proceedings of INLG 2002*, pages 81–88, Harriman, NY, USA, July.

Nikiforos Karamanis. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, Division of Informatics, University of Edinburgh.

Rodger Kibble and Richard Power. 2000. An integrated framework for text planning and pronominalisation. In *Proceedings of INLG 2000*, pages 77–84, Israel.

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

Richard Kittredge, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication knowledge. *Computational Intelligence*, 7:305–314.

Alistair Knott, 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*, chapter 7, pages 181–196. John Benjamins.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, Saporo, Japan, July.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on NLG*, pages 98–107, Niagara-on-the-Lake, Ontario, Canada.

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

Mick O'Donnell, Chris Mellish, Jon Oberlander, and Alistair Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250.

Rebecca J. Passoneau. 1998. Interaction of discourse structure with explicitness of discourse anaphoric phrases. In Walker et al. (Walker et al., 1998b), pages 327–358.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3).

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge.

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

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998a. Centering in naturally occuring discourse: An overview. In Walker et al. (Walker et al., 1998b), pages 1–30.

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