Stochastic Text Structuring using the Principle of Continuity

Nikiforos Karamanis and **Hisar Maruli Manurung**
Institute for Communicating and Collaborative Systems
Division of Informatics
University of Edinburgh
{nikiforo, hisarm}@cogsci.ed.ac.uk

Abstract

This paper explores the feasibility of implementing an evolutionary algorithm for text structuring using the heuristic of *continuity* as a fitness function, chosen over other more complicated metrics of text coherence. Using MCGONAGALL (Manurung et al., 2000) as our experimental platform, we show that by employing an elitist strategy for stochastic search it is possible to quickly reach the global optimum of minimal violations of continuity.

1 Background

Although notions of entity-based coherence have often been employed in text structuring, the definition of an evaluation metric for entity-based coherence is a non-trivial problem. This section reviews some of the suggested solutions and argues for a simpler solution that uses just the principle of continuity as a predictor of the coherence of a text. In the remainder of the paper, we report on our attempt to implement an evolutionary algorithm guided by the heuristic of continuity in order to reach the global optimum quickly and effectively. Finally, we discuss how far this effort stands from actually generating coherent text structures stochastically.

1.1 Text Generation and entity-based models of coherence

The idea of using entity-based constraints on coherence in Natural Language Generation (NLG) goes as far back as McKeown's TEXT generation system (McKeown, 1985). McKeown uses predefined schemata to describe the structure of a text and applies entity-based constraints formulated as local focus rules in order to choose between the alternatives that may match the next predicate in the schema. The proposition that satisfies the most preferred rule for local focus movement is chosen over the rest of the candidates for what to say next.

Subsequent work on NLG tried to move away from predefined schemata by using Rhetorical Structure Theory (RST) as a domain-independent framework for text structuring (Mann and Thompson, 1987). According to RST, a natural text can be described as a hierarchical structure with a rhetorical relation between each two consecutive spans of the text.

More recently, Knott et al. (2001) identified a number of problems in the RST framework concerning the relation OBJECT-ATTRIBUTE ELABORATION. They suggest that ELABORATION be eliminated from the group of Rhetorical Relations and replaced by entity-based models of text coherence such as Centering Theory (Grosz et al., 1995; Walker et al., 1998).

Although Knott et al. (2001) identified Centering Theory (henceforth CT) as one of the possible entity-based models that can be used in the context of text structuring for NLG, the exact formulation of CT in order to serve this purpose remains an open question.

The next section presents the view of Kibble and Power (2000) on how CT can be translated into a part of an evaluation metric that selects the best structure out of a restricted number of candidate solutions. This approach is similar to our view of NLG as a formal search problem, already presented in Mellish et al. (1998). Then, we discuss

how well the metric in Kibble and Power (2000) performs as a predictor of the coherence of texts in a specific domain and argue that a simpler metric based on the principle of continuity appears to yield better results with respect to this task.

1.2 Evaluation metrics for text structuring

Kibble and Power (2000) redefine CT in terms of the "four underlying principles" of entity-based coherence, formally named as *continuity, coherence, cheapness* and *salience.* They describe ICONOCLAST, an NLG system that uses these principles alongside other constraints on text quality.

Although Kibble and Power (2000) mention that each of these principles may be assigned a different cost, in practice they decide that all of them be weighted equally. As a result, their evaluation metric for entity-based coherence is reduced to a function that sums up the number of times that each candidate structure violates each of the underlying principles of CT and then adds the four resulting sums together.

In ICONOCLAST, this metric of entity-based coherence is part of a larger evaluation module that applies a battery of tests to a restricted set of candidate solutions and selects the one with the lowest total cost. Kibble and Power (2000) argue that a candidate solution that violates (some of) the CT-based constraints might still be selected if it respects certain stylistic preferences that are related with the ways of realising the underlying rhetorical structure.

Mellish et al. (1998) were the first to experiment with a range of stochastic search methods in order to select the best rhetorical tree from a number of possible solutions for text structuring. As in Kibble and Power (2000), the evaluation metric in Mellish et al. (1998) includes entity-based features of coherence as well as other parameters of text quality. However, the exact weights that are assigned to the various features of this evaluation metric are based purely on intuition.

Barzilay et al. (2001) present an integrated strategy for ordering information in multidocument summarization. In order to yield a coherent summary, the chronological order of events is combined with a constraint that ensures that sets of sentences on the same topic occur together. This results in a bottom-up approach for ordering that opportunistically groups together topically related sets of sentences.

In this paper, topically related structures are also favoured but since our domain is not predominantly event-based, temporal coherence is not included in our evaluation metric. In the next subsection, we argue that an evaluation function based solely on the principle of continuity represents a simpler and more motivated solution than the ones used by Mellish et al. (1998) and Kibble and Power (2000), at least as far as our genre is concerned.

1.3 The principle of continuity

While both Mellish et al. (1998) and Kibble and Power (2000) investigate the interaction between entity-based coherence and rhetorical relations using intuitive evaluation metrics, Karamanis (2001) follows Knott et al. (2001) in claiming that, in the descriptional genre, text structuring is predominantly entity-based and that rhetorical relations are rare and rather localised. Karamanis (2001) then explores the usefulness of entity-based metrics of text structure in evaluating the overall coherence of a text *without* considering additional constraints such as rhetorical relations. Five evaluation metrics of entity-based coherence are defined and their usefulness as predictors of the coherence in a small corpus of descriptive texts is tested.

The main result is that a simple metric that is based solely on the principle of *continuity*, that is, the requirement that *each utterance in the discourse refers to at least one entity in the utterance that precedes it*,[1] performs better than the other four metrics, including the addition function as defined by Kibble and Power (2000).

Karamanis (2001) uses an input similar to the one we are using in our current experiments.[2] Starting from an "original" ordering of facts that approximates the structure of a descriptive text written by a human expert, all possible orderings are generated by permuting the facts in the original ordering. For each ordering, the total number of violations of the principle of continuity is recorded and compared

[1] A formal definition of this principle in terms of CT is: $Cf(U_{n-1}) \cap Cf(U_n) \neq \emptyset$.

[2] See section 2.1 for more details on the input and the target structures.

with the score of the original ordering which serves as the gold standard. Finally, a complete overview of the number of alternative solutions that score *better*, *equal* or *worse* than the gold standard is obtained.

Karamanis (2001) reports that using a metric that is based solely on the principle of continuity is found to classify on average more than 90% of the search space as *worse* than the original structure. Only 1% of the alternative text structures are found to be *better* than the original one whereas the size of the *equal* solutions is restricted to less than 9% of the search space. Replacing the metric that is based on the principle of continuity with other metrics of entity-based coherence, including the addition function defined by Kibble and Power (2000) consistently gives worse results across the texts in the corpus.[3] More specifically, the average percentages for the Kibble and Power metric are 44% for *better*, 15% for *equal* and only 41% for *worse*.

Ignoring other text structuring factors such as rhetorical relations in the domain of descriptive texts does not prove to be as dangerous as it originally appears, since a metric based on the principle of continuity permits only a limited number of possible orderings to score better than the original structure. Crucially, this metric classifies the original text structure as better than the vast majority of its competitors.

The exhaustive search in Karamanis (2001) revealed a profile of the search space where texts which do not violate continuity are very few indeed. For example, from one text which consists of 12 facts, out of a possible 12! orderings, only 96 orderings (that is, less than 0.0001%) completely satisfied continuity. Furthermore, the orderings that violate continuity once and appear in the same equivalence class as the gold standard represent only 0.0027% of the search space. This suggests that using the principle of continuity to look for the class of optimal texts is a non-trivial search problem.

Due to the factorial complexity of the exhaustive search in Karamanis (2001), the operation becomes impractical for an input that consists of more than 12 facts. In this paper, we extend Karamanis (2001) by discussing a stochastic approach for large inputs which navigates the search space more efficiently. In the section that follows, we provide more details on the methodology that we followed and the software that has been used in order to implement our experiments.

2 Methodology

2.1 Task description

The input to our system consists of an unordered set of facts that correspond to the underlying semantics of a possible text. This input represents the output of the content determination phase of NLG in the standard pipeline architecture. The goal is to find an ordering of all the facts that maximises its entity-based coherence when eventually realised as a text. Although this enforces an artificially rigid distinction between content determination and text structuring, it is necessary for the objective evaluation of the various coherence metrics.[4]

The texts that we are using in our system are short descriptions of archaeological artefacts that have been written in the context of the M-PIRO project (Androutsopoulos et al., 2001). These texts have been analysed into clause-sized propositions so that each clause in the text roughly corresponds to a different proposition in the database.

As a result, a text that originally appears in the surface structure like this:

(1) Towards the end of the archaic period, coins were used for transactions. This particular coin, which comes from that period, is a silver stater from Croton, a Greek Colony in South Italy. On both the obverse and the reverse side there is a tripod (vessel standing on three legs), Apollo's sacred symbol. Dates from between 530-510 BC.

is taken to correspond to the following sequence of facts in the text structure:

[3] The other three metrics of entity-based coherence tested in Karamanis (2001) are (a) a simpler addition function that computes the sum of the violations of only continuity and coherence, (b) a reformulation of that metric in the spirit of Optimality Theory (Prince and Smolensky, 1997) that uses the preference order continuity>coherence, and (c) a similar reformulation of the addition function used by Kibble and Power (2000) that defines the preference order continuity>coherence>cheapness>salience in a way that comes close to some of the predictions of CT. See Karamanis (2001) for more details on the definition and the performance of those metrics.

[4] We follow Kibble and Power (2000), Cheng and Mellish (2000), and Mellish et al. (1998) in this respect.

(2)

1. `use-coins(archaic-period)`
2. `creation-period(ex5, archaic-period)`
3. `madeof(ex5, silver)`
4. `name(ex5, stater)`
5. `origin(ex5, croton)`
6. `concept-description(croton)`
7. `exhibit-depicts({ex5, sides}, tripod)`
8. `concept-description(tripod)`
9. `symbol(tripod, apollo)`
10. `dated(ex5, 530-510bc)`

An unordered set of these facts is the semantic input to our system. The ordering of the facts as defined in example (2) is the target of the text structuring process and serves as the basis for the evaluation of our system.[5] Since we are not concerned with issues of aggregation and realisation of referring expressions, the targeted ordering can be thought to represent a simple surface text as follows:

(3) (1) Towards the end of the archaic period, coins were used for transactions. (2) This coin comes from the archaic period. (3) It is made of silver. (4) It is called a stater. (5) It comes from Croton. (6) Croton is a Greek Colony in South Italy. (7) On both sides of this coin there is a tripod. (8) A tripod is a vessel resting on three legs. (9) It is god Apollo's sacred symbol. (10) The coin dates from between 530-510 BC.

2.2 Evolutionary algorithms

The task of generation does not necessarily require a global optimum (Cheng and Mellish, 2000; Barzilay et al., 2001). What is needed is a text that is *coherent enough* to be understood. Additionally, as stated in Mellish et al. (1998), NLG can benefit from the advantages of an *anytime* algorithm, i.e., an algorithm that can be terminated at any point in time to yield the best result found so far. These two characteristics suggest that the paradigm of **evolutionary algorithms** (henceforth EA) is a good choice for solving our search problem.

[5]The same approach with respect to the target text has been followed by Cheng and Mellish (2000) in their attempt to capture the interaction between aggregation and text structuring by using an evaluation function that extends the one in Mellish et al. (1998).

EAs are a broad class of optimisation methods, to which Genetic Algorithms (GA), employed in both Cheng and Mellish (2000) and Mellish et al. (1998), belong. They are based on a stochastic search process which maintains a population of candidate solutions that evolve according to rules of selection, recombination and mutation. Each candidate receives a measure of fitness in its environment, and selection focuses attention on high fitness individuals. Although simplistic from a biologist's viewpoint, they are sufficiently complex to provide powerful search mechanisms (Spears et al., 1993).

We can characterise our EA with the following algorithm:

$t = 0$
Initialise population $P(t)$ with n random orderings of the given facts.
Evaluate $P(t)$ and rank/select $P(t+1)$
while optimal solution not found or $t <$ maximum iterations **do**
  Evolve $P(t)$ with mutation and/or crossover operations
  Evaluate $P(t)$ and rank/select $P(t+1)$
  $t := t + 1$
end while

Our chosen selection process is the widely-used *roulette-wheel* algorithm, which selects candidate solutions from the previous generation with probability proportional to their fitness values. We also implement an *elitist* strategy, where a small percentage (defined by the *elitist ratio* parameter) of the fittest individuals are always copied over unevolved to the next generation. This guarantees that the best solution found so far is always kept, which improves the EA's overall performance. The trade-off, however, is that it can exert pressure towards premature convergence (Goldberg, 1989).

Fitness Function

Because we require our fitness function to assign a higher score to more continuous texts, we simply count the number of continuity preservations between pairs of subsequent facts. Thus, the theoretical global maximum score achievable given an input semantics of n facts is $n - 1$.

Note, however, that for the `stater` text in section 2.1, even the optimal solutions are bound to violate continuity once, that is, orderings with zero violations of continuity do not exist. So the actual global maximum here is 8. This is still higher than the score of the target structure in (2) which violates

continuity twice (facts `7` and `10`), thus scoring 7.

Operators

- Mutation

  We experimented with three simple mutation operators, i.e. generating a completely random **permutation**, random **swapping** of two facts in an ordering, random **repositioning** of a fact (removing it from its position and inserting it elsewhere, shifting the other facts accordingly).

- Crossover

  We experimented with the combining of subsequences from two orderings x and y by taking a randomly chosen subsequence from x, inserting it at a random point in y, and then removing duplicate facts from the original y. This is how crossover was implemented for the GA experiment in Mellish et al. (1998).

Implementation details

We implemented and ran our experiments using MCGONAGALL, a system being developed with the goal of generating simple rhyme-and-metre poetry, i.e. texts that are highly constrained at both the semantic and surface level (Manurung et al., 2000).

The underlying principles behind this system coincide with the view of NLG as a formal search task, and use EAs to optimise the search. This is motivated by the problems encountered when trying to generate texts with surface constraints by using a traditional semantic goal-driven process.

MCGONAGALL is designed and implemented to be as general-purpose as possible, enabling it to serve as an experimental platform for various evolutionary algorithm-based natural language generation research. Hence, it is an ideal system for our purposes, and conversely, it is hoped that this experiment will test its worth as an experimental platform.

3 Results

Six texts were chosen from the M-PIRO domain, as in section 2.1. Three of these texts contain less than 12 facts, thus the complete profile of their search space is known as a result of the exhaustive search in Karamanis (2001). All experiment results reported in this section are the average results of running the test in question 10 times.

Text name	n facts	Target	Mean	Max.
stater	10	7	6.482	8.0
tetradrachm	10	8	7.602	9.0
drachma	11	9	8.384	10.0
kouros	18	13	14.022	17.0
amphora	20	17	15.328	19.0
hydria	23	20	16.783	20.8

Table 1: Results of the main experiment

Before carrying out our main experiments, we conducted a preliminary experiment to find the most promising choice of parameters for our EA. This was done by running the EA on various possible combinations of choice of operators and elitist ratio parameters. Figure 1 shows the main results of this preliminary experiment.

This figure plots the mean and maximum scores of the population throughout the EA run for 500 iterations on one of the texts in our domain (`amphora`). The horizontal line represents the score of the target structure, i.e. that of the text produced by the human expert. The three columns contrast the results obtained when the *elitist ratio* was varied between 0 (i.e. non-elitist strategy), 0.1, and 0.2. The elitist strategy proved to be crucial in our experiments in guiding evolution towards convergence at an optimal solution without causing serious problems of premature convergence.

The two rows of Figure 1 plot the results of using Crossover and Permute, two of the operators detailed in section 2.2. Here it is shown that Permute performs considerably worse than Crossover. This is because completely random permutation is a highly non-local move in the search space. Swap and Reinsert, the other two operators we experimented with, performed similarly to Crossover.

Table 1 summarises the mean and maximum scores of the population at the end of our main experiment. For these experiments we iterated 4000 times, with a population size of 50, an elitist ratio of 0.2, and we employed the Crossover operator only.

Finally, Figure 2 plots the mean and maximum fitness scores of the population throughout the main experiment for our largest text, `hydria` (23 facts). Similar patterns were found for the other five texts.

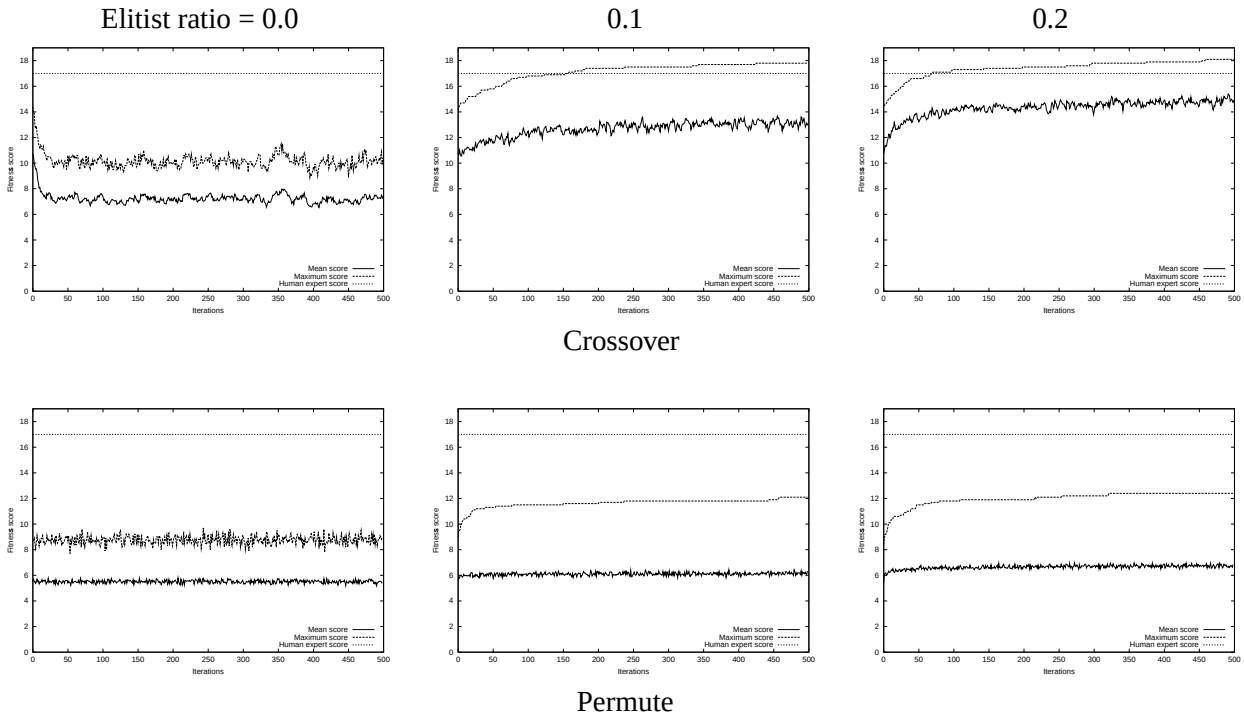

Figure 1: The differences between Crossover and Permute and elitist ratio on `amphora`

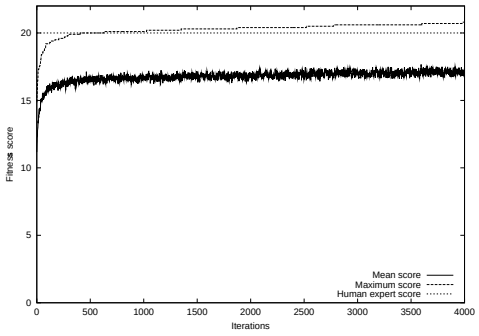

Figure 2: Mean and maximum population scores for `hydria`

4 Discussion

Generally speaking, the graphs show that the population swiftly matches the score of the target structure, and gradually stabilises at a slightly higher score. This is due to the elitist strategy enforcing a hillclimbing-like heuristic within our stochastic process.

Returning to the `stater` text in section 2.1, it is known that the optimal text structure represents 0.05% of a search space of more than 3.6 million alternative orderings. Our EA manages to reach the global maximum of 8 quickly and effectively.

We do not know the percentage of optimal text structures for `hydria`, since with 23 facts it is impractical to profile the vast search space[6] exhaustively. Again, the algorithm reaches a solution close to the target quite quickly.

4.1 Quality of the generated text structures

Although we used the target structure and the profile of the search space as our main basis for evaluating the performance of our system with respect to the search task, we also tried to realise the 10 best structures that the EA produces for the `stater` example as surface texts by hand. As the following example shows most of these texts did not appear to be very different from (3) in section 2.1:

(4) (1) Towards the end of the archaic period, coins were used for transactions. (2) This coin comes from the archaic period. (4) It is called a stater. (3) It is made of silver. (7) On both sides of this coin there is a tripod. (9) The tripod is god Apollo's sacred symbol.

[6] i.e. 2,585,201,673,888,497,664,000 possible orderings!

(8) A tripod is a vessel resting on three legs. (10) The coin dates from between 530-510 BC. (5) It comes from Croton. (6) Croton is a Greek Colony in South Italy.

The main difference between this example and the structure in (3), is that fact 10 in example (4) appears in a position where it satisfies continuity thus reaching the global optimum of only one violation. Note that the original text in (1) avoids the violation of continuity in (3), by aggregating facts 8 and 9 in the same sentence as fact 7.

However, we also noticed that some of the preferred text structures under our approach will actually sound quite incoherent when compared with the original texts. For example, the following text structure starts by focusing on the entity 'tripod', a strategy which does not seem to be preferred in our domain:

(5) (8) A tripod is a vessel resting on three legs. (9) It is god Apollo's sacred symbol (7) On both sides of this coin there is a tripod. (10) The coin dates from between 530-510 BC. (4) It is called a stater. (1) Towards the end of the archaic period, coins were used for transactions. (2) This coin comes from the archaic period. (3) It is made of silver. (5) It comes from Croton. (6) Croton is a Greek Colony in South Italy.

This text also achieves the same score of only one violation, but our metric fails to discriminate between the structure in (4) and the rather incoherent pattern in (5).

5 Future Work

So how far are we from actually generating a coherent text structure using a stochastic approach like the one discussed above? Examples like (5) above suggest that we need to elaborate on our evaluation metric to ensure that we perform stochastic text structuring more effectively.

In the future, we intend to implement a surface generation component based on canned text in order to investigate the output of our experiment more systematically. Spotting continuous text structures that result in incoherent surface texts like the one in (5) allows us to investigate additional principles for text structuring that will supplement continuity and build a more informed evaluation function.

For example, both (4) and (5) are about a coin and not a tripod which might be the reason why (5) is not so good. This example seems to point to the need to incorporate some sort of global coherence into account in the evaluation metric.

Additionally, some of our previous experiments have indicated that not permuting the first utterance in the original sequence might prevent overgeneration, but the results are far from conclusive. We are currently investigating different initialisation strategies in order to prevent structures like the one in (5).

In order to generate a text like the one in (1), we intend to implement an aggregating operator and evaluate its results. This will bring us close to an evaluation function like the one discussed in Cheng and Mellish (2000). Furthermore, we believe that a careful study on the use of the title and the layout in our genre, as well as a better definition of the update unit for local focus might also prevent incoherent structures from being selected.

We also intend to explore the performance of this metric within an integrated architecture that exploits the interaction between content determination and text planning, which MCGONAGALL allows for.

Finally, we recognise that the very limited evaluation set of this experiment might cast doubt on the significance of our results. Therefore, in order to test the generality of our approach we intend to run our experiments on additional texts from the GNOME corpus[7] that have already been annotated semantically in terms of their entity-based coherence and can serve as a suitable input to our experiments.

6 Conclusion

In conclusion, instead of presenting a complete solution for text structuring, the experiments discussed in this paper explore the feasibility of using stochastic search guided by an evaluation function which is based solely on continuity. Keeping this in mind, we have shown that:

- Even though the optimal solutions are quite rare in the search space, a stochastic approach that uses an elitist strategy manages to reach the

[7] http://www.hcrc.ed.ac.uk/~gnome/

global optimum very quickly and avoids premature convergence.

- In most cases, our system generates coherent text structures stochastically by using only the principle of continuity as a fitness function.
- However, some of the resulting surface texts are quite incoherent, and our system gives us the opportunity to investigate the limits of an evaluation metric that is only based on continuity and discuss additional amendments to it.

Acknowledgements

The authors are grateful to Jon Oberlander and Chris Mellish for extended discussions and their general guidance and support, and to three anonymous reviewers for their comments. However, we remain responsible for all the mistakes of the paper. The first author is supported by the Greek State Scholarships Foundation (IKY). The second author is supported by the World Bank QUE Project, Faculty of Computer Science, Universitas Indonesia.

References

Ion Androutsopoulos, Vaki Kokkinaki, Aggeliki Dimitromanolaki, Jonathan Calder, Jon Oberlander, and Elena Not. 2001. Generating multilingual personalized descriptions of museum exhibits: the m-piro project. In *Proceedings of the International Conference on Computer Applications and Quantitative Methods in Archaeology*, Gotland, Sweden.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of HLT*, San Diego.

Hua Cheng and Chris Mellish. 2000. Capturing the interaction between aggregation and text planning in two generation systems. In *Proceedings of INLG-2000*, Israel.

D.E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.

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.

Nikiforos Karamanis. 2001. Exploring entity-based coherence. In *Proceedings of CLUK4*, pages 18–26, University of Sheffield.

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

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. Benjamins, Amsterdam.

William C. Mann 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.

Hisar Maruli Manurung, Graeme Ritchie, and Henry Thompson. 2000. A flexible integrated architecture for generating poetic texts. In *Proceedings of the Fourth Symposium on Natural Language Processing*, Chiang Mai, Thailand, May.

Kathleen R. McKeown. 1985. *Text Generation: Using discourse strategies and focus constraints to generate Natural Language Text*. Studies in Natural Language Processing. Cambridge University Press.

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.

Alan Prince and Paul Smolensky. 1997. Optimality: from neural networks to universal grammar. *Science*, 275:1604–1610.

William M. Spears, Kenneth A. De Jong, Thomas Bck, David B. Fogel, and Hugo de Garis. 1993. An overview of evolutionary computation. In *Machine Learning: ECML-93, Proceedings of the European Conference on Machine Learning*, pages 442–459. Springer.

Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. 1998. Centering in naturally occuring discourse: An overview. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 1–30. Clarendon Press, Oxford.