# A Generative Discourse-New Model for Text Coherence

Micha Elsner and Eugene Charniak

Department of Computer Science
Brown University
Providence, Rhode Island 02912

**CS-07-04**
May 2007

# A Generative Discourse-New Model for Text Coherence

**Micha Elsner and Eugene Charniak**
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{melsner,ec}@cs.brown.edu

## Abstract

Recent models of document coherence have focused on the referents of noun phrases, ignoring their syntax. However, syntax depends on discourse function; NPs which introduce new entities are often more complex. We develop a generative model for NP syntax which describes this difference. It can be used to model discourse coherence in the Wall Street Journal; combining it with the local coherence model of Elsner et al. (2007) yields substantial improvements. Our model is competitive with previous systems on the discourse-new detection task; its performance is comparable to Uryupina (2003).

## 1 Introduction

To a great extent, the syntactic structure of a noun phrase depends on its function within the discourse. If it introduces an unfamiliar (*discourse-new*) entity, it is likely to be marked with some explanatory information:

> Herbert M. Baum, the 53-year-old president of the company's Campbell U.S.A. unit...

*Baum*, for instance, is marked by the use of a full name and an appositional NP explaining who he is. In subsequent mentions within the same text, he appears as merely "Mr. Baum". There is a large class of structures associated in this way with discourse-new NPs[1]: (Poesio and Vieira, 1998) and (Hawkins, 1978) list several, including comparatives, appositives, restrictive relative clauses, and expansions of things (like names) that are usually abbreviated.

We use this variation to model document coherence, a measure of how well the sentences of a document fit together. Coherence models can be used in text generation (Kibble and Power, 2004) and multidocument summarization (Barzilay et al., 2002) to impose an order on the generated or selected sentences. Our work is in the tradition of entity-based coherence models, following (Lapata and Barzilay, 2005; Barzilay and Lapata, 2005). This line of research adapts some generalizations of centering theory (Grosz et al., 1995), tracking references to entities through a discourse and modeling repetition and the assignment of prominent syntactic roles.

To our knowledge, no previous coherence models have exploited discourse-newness. Entity-based models tend to abstract away from the internal syntax of NPs to focus on head nouns or referents. Lexical models (Barzilay and Lee, 2004; Lapata, 2003; Foltz et al., 1998) lack a concept of syntax altogether. Models based on rhetorical relations (Marcu, 2000) tend to operate at the clause level and deal with events.

We construct a generative model of NP syntax conditioned on discourse-newness which can be used (along with a naive heuristic for labelling new NPs) to measure document coherence; thus it ex-

[1]For convenience, we will use *discourse-new NPs* etc. to mean NPs which refer to discourse-new entities. However, this introduces a degree of imprecision; it is important to distinguish synactic categories like NP from semantic and pragmatic categories like newness.

ploits a feature which recent models of coherence have not considered. It can be generatively combined with the relaxed entity grid of Elsner et al. (2007), an extension of Lapata and Barzilay (2005), for a significant increase in performance.

Much work in coreference resolution has focused on detecting discourse-newness, since a discourse-new NP plainly has no referent earlier in the text and need not be resolved by the system. Several studies (Denis and Baldridge, 2007; Poesio et al., 2005; Uryupina, 2003; Ng and Cardie, 2002; Vieira and Poesio, 2000; Bean and Riloff, 1999) have constructed quite accurate discourse-new classifiers[2] and demonstrated their usefulness in coreference resolution. Our model is competitive for the standard discourse-new detection task on the MUC-7 corpus (Hirschman and Chinchor, 1997), a set of news articles from the Wall Street Journal hand-annotated with coreference information.

The linguistics literature has also paid a good deal of attention to discourse-newness and its syntactic marking. Much work has focused on the types of NPs which can be definite, receive focus or be pronominalized (Gundel et al., 1993; Fraurud, 1990; Hawkins, 1978). Fox and Thompson (1990) also provides a detailed explanation of the common discourse uses of relative clauses.

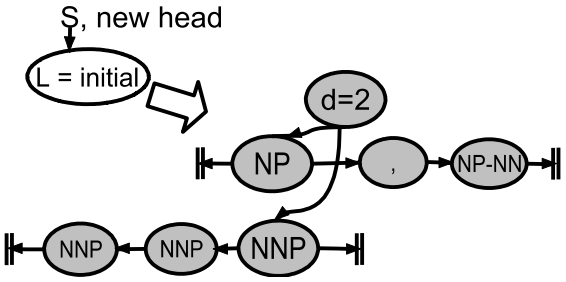


Figure 1: One ($L = initial$) of three possible generations of (NP (NP (NNP Herbert) (NNP M.) (NNP Baum) ) (, ,) (NP the 53-year-old (NN president) (PP of the company 's Campbell U.S.A. unit ))).

## 2 Our Model

Our model is intended to model the syntax of NPs referring to entities; we consider only the largest NP for each head noun. Our grammar is that of the Penn Treebank (M. Marcus et al., 1993), essentially for convenience since we have large quantities of parsed data.

Although the task of discourse-new detection requires only a binary judgement (discourse-new or discourse-old), we follow Fraurud (1990) and divide NPs into three groups; the discourse-new class is split into *initial* and *isolated*[3] . Initial NPs have some coreferent later in the text; isolated NPs do not. Like Fraurud (1990), we find that most NPs are isolated. Discourse-old NPs are rarer, and initial NPs are rarest of all. The three-way distinction is useful to us in our coherence experiments, since reordering a text (treating the coreferential chains as fixed) can make an initial NP into an old one, but not into an isolated NP.

To generate an NP (see figure 1), we first decide its label $L$, *initial*, *discourse-old* or *isolated*, conditioned on the parent non-terminal ($par$) and whether the head noun has occurred in this discourse before ($new$). Using the parent non-terminal allows the model to learn that discourse-new phrases occur more often in some syntactic positions than others (as shown by (Fraurud, 1990)). In the figure, given that the head noun, *Baum*, has not occurred previously, and that the NP's parent is $S$, the label $L$ is chosen as $initial$.

Given $L$, we generate the syntactic structure $S$ in the following way: first we pick $d$, the depth of the substructure. In the figure, $d = 2$, since the phrase has the doubly-nested structure (NP (NP ... ) ). Discourse-new phrases often have more levels than discourse-old since they contain more phrasal modifiers. Given $d$ we generate the actual structure $S_d$ (the head and modifiers) at each nesting level. We generate the lowest level, which contains the head word and its preterminal modifiers, from a different distribution than the others, which contain a head

[2]Bean and Riloff (1999) actually detect non-anaphoric NPs, which include discourse-new NPs and those whose referent is unique, e.g. *the FBI*.

[3]Our discourse-new class is Fraurud (1990)'s "first mention"; our discourse-old class is her "subsequent mention".

phrase and phrasal modifiers. So our model is:

$$P(L,S) = P(L|par,new)P(d|L) \times P(S_d|d,L) \times \prod_{i=1}^{d-1} P(S_i|d,L)$$

The syntax of a level, $S_i$, contains a head $H_i$ and some modifiers to either side, $M_{i,left}$ and $M_{i,right}$. (Head nonterminals are almost always NP, but NX and QP sometimes occur.) The modifiers are generated by a Markov chain, in order from the head outward. The chain terminates by generating a special STOP symbol. In the figure, for instance, $M_{2,left}$ is the set NNP NNP, representing the preterminal modifiers *Herbert M.* to the left of the head *Baum*[4].

$$P(S_i|d,L) = P(H_i|d,L) \times P(M_{i,left}|left,H_i,L) \times P(M_{i,right}|right,H_i,L)$$

$$P(M|dir,h,L) = \prod_{i=1}^{M} P(m_i|m_{i-1,i-2},dir,h,L) \times P(STOP|m_{M,M-1},dir,h,L).$$

Some of the syntactic markers of discourse-newness described by (Poesio and Vieira, 1998) do not have their own specific Penn Treebank nonterminals; for instance one cannot distinguish between different types of relative clause, which are all marked SBAR. Therefore we generate the head preterminals of phrasal modifiers (SBAR-WDT, SBAR-IN, etc.).

Finally, we lexicalize certain types of modifiers. Certain titles or abbreviations are more likely to occur in introductions ("Lt. Greg Geisen", subsequently referred to as "Geisen"), others occur in anaphoric references ("Mr.") and some abbreviations do not alternate at all ("U.S."). We lexicalize any word which is the first or last at its syntactic level and contains a period. In addition, as the long literature on definiteness (Hawkins, 1978; Fraurud, 1990) shows, there are many pragmatic constraints on which entities can be realized by definite or indefinite NPs. Therefore our system distinguishes "the" and "a/an" from other determiners.

[4]As shown in the equations, our model generates the next modifier given the two previous modifiers and the head; the figure shows only a single arrows between adjacent modifiers for legibility.

Estimation of these distributions requires some smoothing. We smooth in a fairly simple manner; we use constant discounting, and for Markov chains we interpolate linearly between estimates of various orders, using the EM algorithm to find interpolation constants.

### 2.1 Bootstrapping

To train our model, we require a large amount of data; when we train the full model on the labeled MUC-7 data, it performs poorly due to sparsity. In an attempt to compensate, we use a bootstrapping scheme; we construct a simplified version of our classifier, which we call the *first stage*, using labeled data, and use it to label a large corpus on which we train the full model (the *bootstrapped* model). Because of sparsity, our first-stage model does not use the full Markov chain to generate modifiers. Instead it distinguishes only two contexts, adjacent to the head and non-adjacent.

On our development set, bootstrapping results in significant gains on the coherence task. However, in testing, our bootstrapped model seems to perform around the same as the first-stage model. Since our coherence test set is much smaller than the development set, we currently guess that the discrepancy is due to chance. However, we report our development numbers as well as final scores.

We expected bootstrapping to improve overall performance on the coherence task because adding more training data sharpens the probability distribution for common features. Thus it may increase the classifier's confidence in its most reliable decisions, and so lead to better overall judgements when many decisions are combined. However, since the training data added is noisy (especially for edge cases), we did not expect bootstrapping to increase the classifier's accuracy.

To bootstrap, we use the first-stage model to classify all NPs in 3 million words of NTC newswire text. We do "soft" bootstrapping; we add each NP to the training data for label $l$ with partial count $P(L = l|NP)$.

We train the first stage on the 30 dryrun documents from the MUC-7 corpus. Following Uryupina (to appear), we segment the data using Reynar and Ratnaparkhi (1997), parse it using Charniak and Johnson (2005) and process all mark-

| | | Prec. | Rec. | F |
|---|---|---|---|---|
| (Ng and Cardie, 2002) | Base | - | - | 73.2 |
| | Res | - | - | 84.0 |
| (Uryupina, to appear) | Base | 66.52 | 100 | 79.89 |
| | Res | 82.29 | 93.54 | 87.56 |
| Generative (Test) | Base | 67.30 | 100 | 80.45 |
| | First Stage | 85.93 | 88.73 | 87.31 |
| | $\Rightarrow$ Bootstrapped $\Leftarrow$ | 85.43 | 87.60 | 86.50 |
| Generative (Development) | Base | 67.95 | 100 | 80.91 |
| | First Stage | 85.54 | 90.44 | 87.92 |
| | Bootstrapped | 85.90 | 91.03 | 88.39 |

Table 1: Results of discourse-new detectors on MUC-7 data.

able NPs. Any NP with an antecedent is labelled discourse-old; the remainder are split between initial and isolated depending on whether they have any coreferents later in the document.

## 3 Modeling Discourse-newness

In discourse-new detection, our goal is simply to classify all markable NPs as discourse-new or discourse-old (again, the discourse-new class is the union of the initial and isolated classes). All previous work on this dataset reports a majority baseline (the scores obtained if all NPs are marked discourse-new), but due to the lack of standardization in parsing the MUC-7 data and extracting markable NPs, there are disagreements in exactly what this baseline should be. For the test set, we obtain 2505 examples of which 1686 are positive. This produces a baseline close to that of (Uryupina, to appear). Our baseline is substantially higher than that of (Ng and Cardie, 2002); we are unable to explain the discrepancy.

As in previous work, we train on the 30 dryrun documents and test on the 20 formal evaluation documents. Development for this task uses crossvalidation on the dryrun documents, treating each document as a separate fold. We report precision, recall and F-score of the discourse-new class. Our results are shown in table 1. The result of the bootstrapped system (marked by $\Rightarrow$ / $\Leftarrow$ arrows) is our final result, selected based on development performance. The remaining results are provided only as a comparison.

These results suggest that our system is competitive with the alternatives[5].

## 4 Modeling Coherence

Only the first item in a coreferential chain can be an initial NP. Thus one measure of the coherence of a document is the likelihood of the syntax of its NPs given that the first item in each chain is initial and the remaining items are discourse-old. Unfortunately, we cannot apply this measure directly. We do not in general know the true coreference relations between NPs, and moreover, if the sentences of the document are out of order, state-of-the-art coreference systems may not work well.

We simplify our approach to overcome this problem by taking all NPs with the same head to be coreferent. This represents a substantial sacrifice; as Poesio and Vieira (1998) show, only about 2/3 of definite descriptions which are anaphoric have the same head as their antecedent. In addition, since we are using the new-head feature to assign labels, we may no longer condition on it in the model itself, so $P(L|par, new)$ becomes $P(L|par)$.

Our model determines the probability of a document $D$ containing a set of NPs. Each NP has a label $L_{np}$ and syntax $S_{np}$. Both the syntax and the labels are visible; we determine the labels using the "coreferential chains" given by our same-head heuristic.

$$P(D) = \prod_{np:NPs} P(L_{np}, S_{np}) = P(L_{np})P(S_{np}|L_{np}). \quad (1)$$

[5]The comparison results we provide represent the state of the art on MUC-7 data. Poesio et al. (2005) report an F-score of 90.2 against a baseline of 67.9 on the GNOME corpus, but have not yet completed tests on MUC-7.

We could use this model to compare one document to another; in this work we only compare different orderings of the same set of sentences. In this case we can make two simplifications: since our input document must contain a fixed set of trees $S_{np}$, we can calculate $P(L_{np}|S_{np}) \propto P(L_{np}, S_{np})$, for which we could in principle use a conditional model rather than a generative one. In addition, reordering the document will not change the status of isolated NPs, so we need not consider them.

### 4.1 Combining with the Entity Grid

Though our model performs fairly well alone ("first-stage" and "bootstrapped" models in table 2), it does not match the performance of previous models. One popular local coherence model is the entity grid (Lapata and Barzilay, 2005; Barzilay and Lapata, 2005). Elsner et al. (2007) describes a "relaxed" variant of this model which applies only to syntactic roles $r$ which will be filled by previously mentioned entities, and predicts for each mentioned entity $e$ the probability $P(r \leftarrow e)$ given a "history" of occurrences for each previously mentioned entity. This model can be generatively combined with (1)[6]; we decide whether an entity will be filled by a mentioned entity (discourse-old) or not, if so, which entity will fill it, and what its syntax will be:

$$P(D) = \prod_{i:NPs} P(L_i)P(r \leftarrow e|L_i = old)P(S_i|L_i). \quad (2)$$

This assumes that the syntax of an NP is independent of which coreferential chain it belongs to, which is untrue (e.g., we expect references to people to have different syntax than references to objects), but could easily be relaxed in future work.

For our "combined" model, the discourse-new component is the bootstrapped model (our best model on development data). We train the entity grid portion of the combined model on sections 2-22 of the Penn Treebank.

### 4.2 Metrics

A popular measure of local coherence has been the discrimination task (Barzilay and Lapata, 2005). In this task, a document is compared with a random permutation of its sentences, and we score the system correct if it indicates the original as more coherent. We use 20 permutations for each document.

Discrimination becomes easier for longer documents, since a random permutation is likely to be much less similar to the original. It is also a somewhat artificial task. Therefore we also test our systems on an additional task, insertion. As motivated by Chen et al. (2007), insertion is a task where an article (e.g. on Wikipedia) is incrementally updated with new information. We use an idealized version of this task to measure the performance of our system. We create a synthetic insertion instance for each sentence of a document, removing it and finding the point of insertion which yields the highest coherence score.

Our insertion metrics are averaged over documents, not sentences, so that longer documents do not have a disproportionate influence on the results. We report two scores: precision is simply the average fraction of sentences per document that are reinserted in their original position. However, this measurement does not distinguish "near misses" from more distant ones, so we also report a positional score; for a document of length $n$, where we have removed sentence $k$ and reinserted it in position $p$ ($1 \leq p, k \leq n$):

$$score = 1 - \frac{|p-k| \times 2n}{k(k-1) + (n-k+1)(n-k)}.$$

This score varies linearly with distance; it takes maximum value 1 for a correct insertion, and expected chance performance is 0. (The minimum diverges slightly from -1 except at the beginning, end, and exact center of the document.)

Both our model and our metrics are "flat"; they ignore the hierarchical structure of the document. Chen et al. (2007) use section and paragraph-level features for insertion and report a hierarchical loss function. We agree that this is a better approach to the problem, but since their model does use local information, we feel that gains in "flat" performance can easily be adapted to improve a hierarchical model as well.

For WSJ, we test on sections 23 and 24 of the Penn Treebank (M. Marcus et al., 1993), 109 documents in total. Our development measurements are

[6] Lapata and Barzilay (2005) use a slightly different model which predicts the role $r$ of each entity. Multiplying this with the discourse-new model does not yield a generative model, since it duplicates part of the prior probability $P(L_i)$.

| | | Discr. (%) | Ins. Prec. (%) | Ins. Pos. |
|---|---|---|---|---|
| WSJ Test | First Stage | 78.07 | 7.73 | 0.0130 |
| | Bootstrapped | 76.65 | 7.82 | 0.0232 |
| | Entity Grid | 85.00 | 10.04 | 0.1692 |
| | $\Rightarrow$ Combined $\Leftarrow$ | 88.62 | 12.35 | 0.2078 |
| WSJ Development | First Stage | 75.33 | 7.44 | 0.0151 |
| | Bootstrapped | 76.85 | 7.52 | 0.0241 |
| | Entity Grid | 83.14 | 10.23 | 0.1690 |
| | Combined | 86.87 | 13.34 | 0.1890 |

Table 2: Results of local coherence models on WSJ.

crossvalidated on sections 2-22, using each section as a fold, for 1199 documents in total. We discard documents with fewer than 8 sentences.

### 4.3 Results

In both test and development, the combined system is significantly better ($p < .05$ using a one-tailed Fisher sign test) than the entity grid alone on each of the three metrics. On test, the bootstrapped model does not differ significantly from the first-stage model at this level on any of the metrics. On our development set, which is much larger than the test set, we find it improves significantly for discrimination and for the positional insertion metric.

Our results (table 2) are the first reported for these coherence metrics on WSJ data. The insertion task is new; previous work on the discrimination task reported results only for the AIRPLANE corpus (Barzilay and Lapata, 2005), which contains short press releases on airplane crashes. The "relaxed" entity grid of Elsner et al. (2007) performs reasonably on that corpus, and we feel it is an appropriate baseline for WSJ.

AIRPLANE texts were originally written for a highly specific set of readers who are experts on airplanes, aeronautical engineering and airfield locations. In addition, they are composed in a very constrained and formulaic style. They do not use prolonged introductions for even the most technical information, and contain little syntactic variety. Our combined model yields a slight improvement on AIRPLANE development data, but the corpus does not contain enough useful syntactic features to allow much improvement.

The WSJ corpus is much more appropriate for our technique. It contains news articles designed for a general audience. These texts introduce a variety of unfamiliar entities (including people, companies, objects and abstract concepts), and are written in a natural style using a variety of syntactic constructions. Most English texts are more similar to WSJ than AIRPLANE; we feel that developing coherence models to improve performance for WSJ will produce more robust and general systems[7].

### 4.4 Syntactic Marking and Hearer-newness

An analysis of the flaws in our algorithm reveals some interesting patterns. Our use of the discourse-new detector for coherence modeling assumes that the detector will work equally well on all discourse-new nouns (even without the new-head feature). This turns out not to be the case. As previous work has recognized, discourse-new detectors must place a great deal of weight on the new-head feature or a more sophisticated search for some likely antecedent earlier in the text (Vieira and Poesio, 2000; Ng and Cardie, 2002). This is because many discourse-new NPs are syntactically unmarked.

Syntactic marking is not spread uniformally over the space of NPs, but is attached to a particular semantic class. The hearer-new category (Prince, 1992) is the set of entities which the speaker expects to be unknown (or unfamiliar) to the hearer. Not all discourse-new entities are hearer-new: some are known to the hearer in advance (the *unused* (Prince, 1981) or *unique* (Uryupina, 2003) class: *the United States*, *Thursday*). Others, Prince's *inferrable* class, are obvious from context ("I walked

[7]The corpus used for our discourse-new detection experiments, MUC-7, also contains Wall Street Journal articles, but is much smaller than WSJ. The results obtained are broadly similar, but less reliable.

home and opened *the door*"). These entities do not usually receive the special modifiers associated with discourse-newness, and it is pragmatically strange to apply them: "*Thursday, the day right after Wednesday*", "I walked home and opened *the door which leads into my house*". Such phrases tend to seem sarcastic or condescending.

Luckily, not every chain whose first element is unmarked degrades our performance. Many such NPs are isolated and can be ignored for the purpose of selecting an ordering. For the same reason, mislabeling the first item in a sequence of NPs with a unique referent ("Thursday") will not always cause an error, since all the NPs in such a chain usually have similar (or identical) structures, and in this case the model does not prefer any particular ordering of them. However, if the NPs do have small differences in structure, errors can occur; we detected several which reduced our performance on the development set.

## 5 Future Work

The systems described here contain a variety of unwarranted assumptions, and so could presumably be improved in a variety of ways. The most obvious is probably also the most ambitious; our coherence model would work much better if we detected coreferential chains using a probabilistic model of NP coreference rather than the "same head" heuristic.

As discussed, our coherence model also makes some errors due to our assumption that all discourse-new NPs will be distinguished by their syntactic form. If we could reliably distinguish unique and inferrable NPs from hearer-new NPs, our model could more accurately describe their syntax. Some previous work has focused on each of these classes: for uniques, Uryupina (2003) demonstrates a relatively accurate supervised classifier, and an unsupervised detector is part of (Bean and Riloff, 1999). Poesio et al. (1997) uses Wordnet to attempt to detect inferrables. Berland and Charniak (1999) use a pattern-matching heuristic to identify "parts" of objects, where "parts" include qualities as well as physical components; many such parts seem to be inferrables. Another approach to finding uniques and inferrables would be to use our model directly. For documents with known coreferential chains, we can simply select all the discourse-new NPs which our syntactic model cannot classify without the new-head feature. According to our hypothesis, these mistakes are likely to be uniques or inferrables.

In the field of coherence modeling, a variety of models make use of global document structure, using hierarchical features or non-local dependencies: (Chen et al., 2007; Elsner et al., 2007; Soricut and Marcu, 2006; Barzilay and Lee, 2004). This work has shown that combining local and global features invariably improves performance. Thus, discourse-new information should be added to the list of local features which are used in such model combinations.

## 6 Acknowledgements

This tech report was prepared with the generous assistance of Olga Uryupina, Regina Barzilay and Erdong Chen, and discussed extensively with Joe Austerweil, Matt Lease, David McClosky, Jenine Turner, and Greg Shakhnarovich. We also thank Dan Jurafsky for references, Keith Hall for data, and four anonymous EMNLP reviewers for their comments. The authors were supported by DARPA GALE contract HR0011-06-2-0001.

## References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

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

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

David L. Bean and Ellen Riloff. 1999. Corpus-based identification of non-anaphoric noun phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL'99)*, pages 373–380, Morristown, NJ, USA. Association for Computational Linguistics.

Mathew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Orlando, Florida. Harcourt Brace.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proc. of the 2005 Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 173–180.

Erdong Chen, Benjamin Snyder, and Regina Barzilay. 2007. Incremental text structuring with online hierarchical ranking. In *Proceedings of EMNLP*.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester, New York, April. Association for Computational Linguistics.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of HLT-NAACL '07, to appear*.

Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

Barbara Fox and Sandra A. Thompson. 1990. A discourse explanation of the grammar of relative clauses in English conversation. *Language*, 66(2):297–316.

Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395–433.

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.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

John A. Hawkins. 1978. *Definiteness and indefiniteness: a study in reference and grammaticality prediction*. Croom Helm Ltd.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 Coreference Task definition. In *Message Understanding Conference Proceedings*.

Roger Kibble and Richard Power. 2004. Optimising referential coherence in text generation. *Computational Linguistics*, 30(4):401–416.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the annual meeting of ACL, 2003*.

M. Marcus et al. 1993. Building a large annotated corpus of English: The Penn Treebank. *Comp. Linguistics*, 19(2):313–330.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging descriptions in unrestricted text. *ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts*, pages 1–6.

Massimo Poesio, Mijail Alexandrov-Kabadjov, Renata Vieira, Rodrigo Goulart, and Olga Uryupina. 2005. Does discourse-new detection help definite description resolution? In *Proceedings of the Sixth International Workshop on Computational Semantics*, Tillburg.

Ellen Prince. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.

Ellen Prince. 1992. The ZPG letter: subjects, definiteness, and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text,*, pages 295–325, Philadelphia/Amsterdam. John Benjamins B.V.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington D.C.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the Association for Computational Linguistics Conference (ACL-2006)*.

Olga Uryupina. 2003. High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*, Sapporo.

Olga Uryupina. to appear. *Knowledge acquisition for coreference resolution*. Ph.D. thesis, University of Saarland.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.