Automated Evaluation of Coherence in Student Essays

Eleni Miltsakaki*, **Karen Kukich**†

*University of Pennsylvania
619 Williams Hall, 36th & Spruce St., Philadelphia, PA 19104-6305, U.S.A.
elenimi@unagi.cis.upenn.edu

†Educational Testing Service
Rosedale Road, Princeton, NJ 08541, U.S.A.
kkukich@ets.org

Abstract

This paper explores the role of Centering Theory, in particular Rough-Shift identification, in locating abrupt topic shifts in student essays. Rough-Shifts within student paragraphs are generated by short-lived topics and are therefore indicative of poor topic development. We develop a Rough-Shift-based metric of incoherence to represent a coherence criterion in essay evaluation. We show that adding this metric to an existing system for automated writing evaluation, ETS's *e-rater*, improves its performance significantly, better approximating human scores and providing the capability for valuable instructional feedback to the student.

1. Introduction

The task of evaluating a student's writing ability has traditionally been a labor-intensive human endeavor. However, several different software systems, e.g., PEG (Page and Peterson, 1995), Intelligent Essay Assessor [1] and *e-rater* [2], are now being used to perform this task fully automatically. Furthermore, by at least one measure, these software systems evaluate student essays with the same degree of accuracy as human experts. That is, computer-generated scores tend to match human expert scores as frequently as two human scores match each other.

In this study, we exploit the availability of such software systems and rich resources of electronically available student essays to test a theoretical hypothesis derived from the Centering Model of discourse coherence (Joshi and Weinstein, 1981; Grosz et al., 1983, inter alia). We propose a metric of incoherence based on the relative proportion of Centering Rough-Shift transitions. We employ the *e-rater* scoring system to test the hypothesis that the Rough-Shift metric is a significant contributor to the accuracy of computer-generated essay scores. Our positive finding suggests a route for exploring Centering Theory's practical applicability to writing evaluation and instruction.

2. The *e-rater* essay scoring system

Approaches to essay scoring vary in their use of NLP techniques and other methods to assess the writing ability exhibited in an essay. Very early work by Page (1966), Page (1968), Page and Peterson (1995) demonstrated that computing the fourth root of the number of words in an essay provides a highly accurate technique for predicting human-generated essay scores. Such measures of essay length have two main weaknesses which render them impractical for writing evaluation. First, scoring criteria based on a superficial word count make the automated system susceptible to deception. Furthermore, due to their lack of explanatory power, such measures cannot be translated into instructional feedback to the student. To improve the efficiency of automated writing evaluation systems, we need to build models which more closely represent the criteria that human experts use to evaluate essays.

Two more recent approaches have attempted to define computational techniques based on these criteria. Both of these approaches are able to predict human scores with at least as much accuracy as length-based approaches. One of these systems, the Intelligent Essay Assessor (Landauer, 1998; Foltz et al., 1998; Schreiner et al., 1997), employs a technique called Latent Semantic Analysis (Deerwester et al., 1990) as a measure of the degree to which the vocabulary patterns found in an essay reflect the writer's semantic and linguistic competence. Another system, the Electronic Essay Rater, *e-rater*, (Burstein et al., 1998), employs a variety of NLP techniques, including sentence parsing, discourse structure evaluation, and vocabulary assessment techniques to derive values for over fifty writing features.

The writing features that *e-rater* evaluates were specifically chosen to reflect scoring criteria defined by ETS writing evaluation experts for the essay portion of the Graduate Management Admissions Test (GMAT). These criteria are fully articulated in GMAT test preparation and scoring materials, which can be found at http://www.gmat.org. Based on these criteria, syntactic variety is represented by features that quantify occurrences of clause types. Logical organization and clear transitions are represented by features that quantify cue words in certain syntactic constructions. The existence of main and supporting points is represented by features that detect where new points begin and where they are developed. *E-rater* also includes features that quantify the appropriateness of the vocabulary content of an essay.

One feature of writing valued by writing experts that is not explicitly represented in the current version of *e-rater* is coherence. Centering Theory provides an algorithm for computing local coherence in written discourse. Our study investigates the applicability of Centering Theory's local coherence measure to essay evaluation by determining the effect of adding this new feature to *e-rater*'s existing array

[1] http://lsa.colorado.edu.
[2] http://www.ets.org/research/erater.html

of features.

3. The Centering model

Centering Theory models the local focusing level of attentional state in discourse and is intended as a component of a theory of local discourse coherence (Joshi and Kuhn, 1979; Joshi and Weinstein, 1981; Sidner, 1979; Grosz, 1977; Grosz and Sidner, 1986). According to Centering, discourse consists of a sequence of textual segments and each segment consists of a sequence of utterances designated by $U_i - U_n$. Each utterance U_i evokes a *set* of discourse entities, the FORWARD-LOOKING CENTERS, designated by $Cf(U_i)$. The members of the Cf set are ranked according to discourse salience (the ranking rule is given in 3.2). The highest-ranked member of the Cf set is the PREFERRED CENTER, Cp. A BACKWARD-LOOKING CENTER, Cb, is also identified for utterance U_i.The highest ranked entity in $Cf(U_{i-1})$ *realized* in U_i is called the BACKWARD-LOOKING CENTER, Cb. (If an utterance is segment initial, then it has no Cb.) The Cb is a special member of the Cf set which can be best understood as what in the literature is often called the 'topic' or 'focus' (Reinhart, 1981; Horn, 1986). Perceived topic shifts are the result of establishing new Cbs.

The Cp for a given utterance may be identical with its Cb, but not necessarily so. The Cb, the BACKWARD-LOOKING CENTER, is a link from the current utterance to the previous discourse. The PREFERRED CENTER is a prediction about the Cb of the following utterance. This distinction between looking back in the discourse with the Cb and projecting preferences for interpretations in the subsequent discourse with the Cp is the key element in computing local coherence in discourse.

3.1. Centering transitions

Four types of transitions, reflecting four degrees of coherence, are defined in Centering. They are computed as shown in Table 1 and ordered according to the ordering rule in (1).

	Cb(Ui) = Cb(Ui-1)	Cb(Ui) $\neq$ Cb(Ui-1)
Cb(Ui) = Cp	Continue	Smooth-Shift
Cb(Ui) $\neq$ Cp	Retain	Rough-Shift

Table 1: Table of transitions

(1) Transition ordering rule:
Continue is preferred to Retain, which is preferred to Smooth-Shift, which is preferred to Rough-Shift.

3.2. Cf ranking.

The ordering of the Cf list plays a crucial role in determining the type of transition holding between two consecutive utterances. The salience status of an entity may be determined by a number of factors. Kameyama (1985) and Brennan et al. (1987) proposed that the Cf ranking for English is determined by grammatical function as follows:

(2) Rule for the ranking of FORWARD-LOOKING CENTERS:
SUBJECT>IND. OBJECT>OBJECT>OTHERS

Later cross-linguistic studies based on empirical work (Di Eugenio, 1998; Turan, 1995; Kameyama, 1985) further refined the ranking, shown in (3), with QIS standing for quantified indefinite subjects (*people*, *one* etc) and PRO-ARB for arbitrary plural pronominals (*we*, *you* in the generic sense).

(3) Revised rule for the ranking of FORWARD-LOOKING CENTERS:
SUBJECT>IND. OBJECT>OBJECT>OTHERS>QIS, PRO-ARB

4. The *e-rater* Centering study

In an earlier preliminary study, we applied the Centering algorithm manually to a sample of 36 GMAT essays to explore the hypothesis that the Centering model provides a reasonable measure of coherence (or lack of), reflecting the evaluation performed by GMAT scoring experts. These experts were trained according to the criteria in the GMAT scoring guide. We observed that essays with higher scores (5-6) tended to have significantly lower percentages of ROUGH-SHIFTs than essays with lower scores. As expected, the distribution of the other types of transitions was not significant. In general, CONTINUEs, RETAINs, and SMOOTH-SHIFTs do not yield incoherent discourses (in fact, an essay with only CONTINUE transitions might sound rather boring!).

In this study we test the hypothesis that a predictor variable derived from Centering can significantly improve the performance of *e-rater*. Since we are in fact proposing Centering's ROUGH-SHIFTs as a predictor variable, our model, strictly speaking, measures *incoherence*.

The corpus for our study came from a pool of essays written by students taking the GMAT test. We randomly selected a total of 100 essays, covering the full range of the scoring scale, where 1 is lowest and 6 is highest (see appended Table 4). We applied the Centering algorithm to all 100 essays, calculated the percentage of ROUGH-SHIFTs in each essay and then ran multiple regression to evaluate the contribution of the proposed variable to *e-rater*'s performance.

4.1. Centering assumptions and modifications

Utterance. In an earlier formulation of Centering the 'utterance' was not defined explicitly. In subsequent work, (Kameyama, 1998), the utterance was defined as, roughly, the tensed clause with relative clauses and clausal complements as exceptions. Recent cross-linguistic studies lead to another revision according to which the utterance is defined as the traditional 'sentence', i.e., the main clause and its accompanying subordinate and adjunct clauses constitute a single utterance (Miltsakaki, 1999). Here, we adopt Miltsakaki's revised definition.

Cf ranking. We assumed the Cf ranking given in (3). A modification we made involved the status of the pronominal

I.[3] We observed that in poor essays the first person pronominal *I* was used extensively, normally presenting personal narratives. However, personal narratives were unsuited to this essay writing task. The extensive use of *I* in the subject position produced an unwanted effect of high coherence. We prescriptively decided to penalize the use of *I*'s in order to better reflect the coherence demands made by the particular writing task. The way to penalize was to omit *I*'s. As a result, coherence was measured with respect to the treatment of the remaining entities in the *I*-containing utterances. This gave us the desired result of being able to distinguish those *I*-containing utterances which made coherent transitions with respect to the entities they were talking about and those that did not.

Segments. Segment boundaries are extremely hard to identify in an accurate and principled way. Furthermore, existing algorithms (Morris and Hirst, 1991; Youmans, 1991; Hearst, 1994; Kozima, 1993; Reynar, 1994; Passonneau and Litman, 1997; Passonneau, 1998) rely heavily on the assumption of textual coherence. In our case, textual coherence cannot be assumed. Given that text organization is also part of the evaluation of the essays, we decided to use the students' paragraph breaks to locate segment boundaries.

4.2. Implementation

For this study, we decided to manually tag coreferring expressions despite the availability of coreference algorithms. We made this decision because a poor performance of the coreference algorithm would give us distorted results and we would not be able to test our hypothesis. For the same reason, we manually tagged the PREFERRED CENTERS as Cp. We only needed to mark all the other entities as OTHER. This information was adequate for the computation of the transitions.

Discourse segmentation and the implementation of the Centering algorithm for the computation of the transitions were automated. Segments boundaries were marked at paragraph breaks and the transitions were calculated according to the instructions given in Table 1. In the output the system gave the percentage of Rough-Shifts for each essay. The percentage of Rough-Shifts was calculated as the number of Rough-Shifts over the total number of identified transitions in the essay.

4.3. An example of *coherent* text

What follows is a small excerpt (a paragraph) of an student essay scored 6.[4] For each utterance, enclosed in the <UT> and </UT> tags, the PREFERRED CENTER and OTHER entities are tagged as <CP> and <OTHER> respectively. Each entity is assigned a unique ID number, REF. Following each utterance, the Cb, Cp and transition type are identified. The following paragraph demonstrates an example of a maximally coherent text, centering the company 'Famous name's Baby Food' and continuing with the same center through the entire paragraph.

<UT> Yet another company that strives for the "big bucks" through conventional thinking is <CP REF='3'>Famous name's Baby Food</CP>.</UT> Cb=none Cp=3 Tr=none
<UT><CP REF='3'>This company</CP> does not go beyond the norm in their product line, product packaging or advertising.</UT> Cb=3 Cp=3 Tr=Continue
<UT>If they opted for an extreme market-place, <CP REF='3'>they</CP> would be ousted.</UT> Cb=3 Cp=3 Tr=Continue
<UT>Just look who <CP REF='3'>their</CP> market is!</UT> Cb=3 Cp=3 Tr=Continue
<UT>As new parents, <CP REF='3'>the Famous name</CP> customer wants tradition, quality and trust in their product of choice.</UT> Cb=3 Cp=3 Tr=Continue
<UT><CP REF='3'>Famous name</CP> knows this and gives it to them by focusing on "all natural" ingredients, packaging that shows the happiest baby in the world and feel good commercials the exude great family values.</UT> Cb=3 Cp=3 Tr=Continue
<UT><CP REF='3'>Famous name</CP> has really stuck to the typical ways of doing things and in return has been awarded with a healthy bottom line.</UT> Cb=3 Cp=3 Tr=Continue

4.4. An example of *incoherent* text

Following the same mark-up conventions, we demonstrate text incoherence with an excerpt (a paragraph again) of a student essay scored 4. In this case, repeated Rough-Shift transitions are identified. Several entities are centered,*opinion*, *success* and *conventional practices*, none of which is linked to the previous or following discourse. This discontinuity created by the very short lived Cbs makes it hard to identify the topic of this paragraph and at the same time it is capturing the fact that the introduced centers are poorly developed.

<UT>I disagree with <CP REF='1'>the opinion</CP> stated above.</UT> Cb=none Cp=1 Tr=none
<UT>In order to achieve <CP REF='4'>real and lasting success</CP> <OTHER REF='2'>a person</OTHER> does not have to be a billonaire.</UT> Cb=none Cp=4 Tr=Rough-Shift
<UT>And also because <CP REF='3'>conventional practices and ways of thinking</CP> can help a person to become rich.</UT> Cb=2 Cp=3 Tr=Rough-Shift

5. Study results

In the appended Table 4, we give the percentages of Rough-Shifts (ROUGH) for each of the actual student essays (100) on which we tested the ROUGH variable in the regression discussed below. The HUM column contains the corresponding scores given by human raters and the E-R column contains the corresponding score assigned by

[3]In fact, a similar modification was proposed by Hurewitz (1998) and Walker (1998) who observed that the use of *I* in sentences such as 'I believe that...', 'I think that...' does not affect the focus structure of the text.

[4]Only proper names have been changed for privacy protection. Spelling and other typographical errors have been corrected, also for privacy reasons.

e-rater. Comparing HUM and ROUGH, we observe that essays with scores from the higher end of the scale tend to have lower percentages of Rough-Shifts than the ones from the lower end. To evaluate if this observation can be utilized to improve the *e-rater*'s performance, we regressed X=E-RATER and X=ROUGH (the predictors) by Y=HUMAN. The results of the regression are shown in Table 3. In evaluating the contribution of a single variable, the t-test is sufficient for testing the null hypothesis. The t-test for the contribution of ROUGH (shown in the 'Parameters Estimates' section of Table 2) has a highly significant p value ($p<0.0013$) indicating that the null hypothesis should be rejected and that including the tested variable significantly contributes to the estimation of the predicted values. In our case, this means that adding ROUGH to E-RATER improves the accuracy of the predicted values. For ease of comparison, Table 2 shows the results of the regression run on E-RATER alone. Using E-RATER as the sole variable yields smaller coverage of the data, shown in the RSquare value, and higher error, shown in the 'Lack of Fit' section of the table.

In evaluating the magnitude of the effect of the ROUGH variable we observe that the ROUGH coefficient modifies the *e-rater*'s scores by approximately .5 point, a reasonably sizeable effect given the scoring scale. Table 4, also, contains the predicted values generated by E-RATER (PrH/E) as the sole variable in the model and the predicted values generated by including both E-RATER and ROUGH (PrH/E+R) in the model. We observe that the predicted values with E-RATER and ROUGH are tilting the scores in the right direction, better approximating the HUM(AN) scores. In particular, in discrepant cases (where *e-rater*'s score differs by more than one point from the human score), the PrH/E+R value makes improvements in all 8 out of the 8 such cases in this data set. In the cases where *e-rater*'s score is adjacent to the human score (different by 1 point in either direction), PrH/E+R makes improvements in 31 out of the 48 cases with the best performance in the high scored essays and the poorest in the essays scored 1 and 2. This is not surprising because essays scored 1 and 2 are normally couple of lines long with minimal to zero number of transitions.

6. Discussion

Our positive finding, namely that Centering Theory's measure of relative proportion of Rough-Shift transitions is indeed a significant contributor to the accuracy of computer-generated essay scores, has several practical and theoretical implications. Clearly, it indicates that adding a local coherence feature to *e-rater* could significantly improve *e-rater*'s scoring accuracy. More important for teaching and evaluation of writing, however, is the fact that the Rough-Shift algorithm provides students and teachers with pointers to sections of their essays where Rough-Shifts occur. Such a feature would add an instructional capability to *e-rater* by providing valuable explanatory power. That is, in addition to an overall essay score, the system could also generate a coherence score for an essay. It could even highlight specific portions of text within an essay where Rough Shifts occur. This information can then be discussed by students with their instructors to gain insight in how to alter the text in the area where Rough-Shifts occurred to improve the local coherence of the essay. Also, note that overall scores and coherence scores need not be strongly correlated. In our data, we find several examples where coherence scores were higher than the overall score and vice versa.

We briefly reviewed these cases with several ETS writing assessment experts to gain their insights into the value of pursuing this work further. In an effort to maximize the use of their time with us, we carefully selected three pairs of essays to elicit specific information. One pair included two high-scoring (6) essays, one with a high coherence score and the other with a low coherence score. Another pair included two essays with low coherence scores but differing overall scores (a 5 and a 6). A final pair was carefully chosen to include one essay with an overall score of 3 that made several main points but did not develop them fully or coherently, and another essay with an overall score of 4 that made only one main point but did develop it fully and coherently.

After briefly describing the Rough-Shift coherence measure and without revealing either the overall scores or the coherence scores of the essay pairs, we asked our experts for their comments on the overall scores and coherence of the essays. In all cases, our experts precisely identified the scores the essays had been given. In the first case, they agreed with the high Centering coherence measure, but one expert disagreed with the low Centering coherence measure. For that essay, one expert noted that "coherence comes and goes" while another found coherence in a "chronological organization of examples" (a notion beyond the domain of Centering Theory). In the second case, our experts' judgments confirmed the Rough-Shift coherence measure. In the third case, our experts specifically identified both the coherence and the development aspects as determinants of the essays' scores. In general, our experts felt that the development of an automated coherence measure would be a useful instructional aid.

7. Remaining issues

The Rough-Shift algorithm relies heavily on the efficiency of automated coreference systems. Discourse deictic expressions and nominalizations are especially hard for such systems and raise a number of interesting research projects. We discuss these issues below.

Discourse deixis describes the phenomenon whereby speakers use demonstrative expressions such as 'this' and 'that' to refer to propositions or in general lengthier parts of the preceding discourse. Webber (1991) argued that referents for discourse deixis are provided by discourse segments on the right frontier of a formal tree structure. However, what the status of such entities is within the Centering framework remains unclear. A possible future direction would be to conduct psychological experiments to test the effect that the use of such expressions has on the processing load imposed on the speaker, compared with simpler entities such as *John* or *the newspaper*. [5]

[5] It seemed to us that the judgments required to establish even a working hypothesis were too fine to make and so we decided to omit the utterances including discourse deictic expressions.

Summary of Fit				
RSquare	0.69733			
Root Mean Square Error	0.876943			
Lack of Fit				
Source	Mean Square	F Ratio		
Lack of Fit	0.950663	1.2521		
Pure Error	0.759263	Prob>F		
Total Error		0.2914		
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	0.4860303	0.243538	2.00	0.0487
E-RATER	0.9177338	0.061076	15.03	<.0001
Effect Test				
Source Nparm	DF	Sum of Squares	F Ratio	Prob>F
E-RATER 1	1	173.63523	225.7853	<.0001

Table 2: Regression on ERATER

Summary of Fit				
RSquare	0.724403			
Root Mean Square Error	0.8422			
Lack of Fit				
Source	Mean Square	F Ratio		
Lack of Fit	0.754228	1.3085		
Pure Error	0.576389	Prob>F		
Total Error		0.2336		
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> $\|t\|$
Intercept	1.4676859	0.374668	3.92	0.0002
E-RATER	0.8053165	0.067634	11.91	<.0001
ROUGH	-0.01393	0.004197	-3.32	0.0013
Effect Test				
Source Nparm	DF	Sum of Squares	F Ratio	Prob>F
E-RATER 1	1	100.56075	141.7746	<.0001
ROUGH 1	1	7.81476	11.0176	0.0013

Table 3: Regression on ERATER and ROUGH

In addition to discourse deixis, the status of nominalizations of verbs or verb phrases is also unclear. The issue of nominalizations (essentially, another form of discourse deixis) raises itself in cases where a coherence link could arguably be established between the verb of one utterance and a nominalized version of it, occurring in the subsequent utterance. To give an example, it is possible that in (1) and (2) below the coherence link is established by the semantics of the verb 'changes' and the noun 'change'.

(1) Many software companies changed their policy.

(2) This change brought about a series of new problems.

One problem in integrating this intuition into the current model is that it is not obvious how we should represent verb meanings in the Cf set and what the relevant ranking of such entities would be. Also, even if we forced our system to detect these cases by comparing the verbs and nouns on a lexicomorphological level we would still miss cases where the link is based on synonymity or more complex inferencing. Since this issue remains unsolved, those potential links were simply missed by our system. Fortunately, such cases were rare. In our corpus, there were only three such instances.

8. Future work

Our study prescribes a route for several future research projects. A full study will require a larger corpus of student essays. Additional programming will be required to implement the Rough-Shift identification algorithm. Research in coreference resolution, discourse deixis and nominalizations is essential for converting this measure to a complete and fully automated procedure. Consulting with writing experts will also be necessary for the evaluation and weight of the Rough-Shift metric of incoherence.

Acknowledgements

We thank several individuals for their valuable contributions to this work. Jill Burstein provided us with the essay set and corresponding human and *e-rater* scores used in this study. Mary Fowles, Peter Cooper, and Seth Weiner provided us with the valuable insights of their writing assessment expertise. Ramin Hemat provided us with the perl code for automatically computing Centering transitions and the Rough-Shift measure for each essay. We are grateful to Aravind Joshi and Alistair Knott for useful discussions.

9. References

S. Brennan, M. Walker-Friedman, and C. Pollard. 1987. A Centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162. Stanford, Calif.

J. Burstein, K. Kukich, S. Wolff, M. Chodorow, L. Braden-Harder, M.D. Harris, and C. Lu. 1998. Automated essay scoring using a hybrid feature identification technique. In *Annual Meeting of the Association for Computational Linguistics, Montreal, Canada*, August.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

B. Di Eugenio. 1998. Centering in Italian. In *Centering Theory in Discourse*, pages 115–137. Clarendon Press, Oxford.

P. Foltz, W. Kinstch, and T. Landauer. 1998. The measuerement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307.

B. Grosz and C. Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

B. Grosz, A. Joshi, and S. Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Annual Meeting of the Association for Computational Linguistics*, pages 44–50.

B. Grosz. 1977. The representation and use of focus in language underastanding. Technical Report No. 151, Menlo Park, Calif., SRI International.

M. Hearst. 1994. Multiparagraph segmentation of expository text. In *Proc. of the 32nd ACL.*

L. Horn. 1986. Presupposition, theme and variations. In *Chicago Linguistics Society*, volume 22, pages 168–192.

F. Hurewitz. 1998. A quantitative look at discourse coherence. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, chapter 14. Clarendon Press, Oxford.

A. Joshi and S. Kuhn. 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *6th International Joint Conference on Artificial Intelligence*, pages 435–439.

A. Joshi and S. Weinstein. 1981. Control of inference: Role of some aspects of discourse structure: centering. In *7th International Joint Conference on Artificial Intelligence*, pages 385–387.

M. Kameyama. 1985. *Zero Anaphora: The Case of Japanese*. Ph.D. thesis, Stanford University.

M. Kameyama. 1998. Intrasentential Centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Clarendon Press: Oxford.

H. Kozima. 1993. Text segmentation based on similarity between words. In *Proc. of the 31st ACL (Student Session)*, pages 286–288.

T. Landauer. 1998. Introduction to latent semantic analysis. *Discourse Processes*, pages 259–284.

E. Miltsakaki. 1999. Dissociating discourse salience from information structure: Evidence from a centering study in Modern Greek and Japanese. In *Computational Linguistics in the Netherelands, CLIN '99*.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of the text. *Computational Linguistics*, 17:21–28.

E. B. Page and N. Peterson. 1995. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, March:561–565.

E. Page. 1966. The imminence of grading essays by computer. In *Phi Delta Kappan*, volume 48, pages 238–243.

E. Page. 1968. Analyzing student essays by computer. *International Review of Education*, 14:210–225.

R. Passonneau and D. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

R. Passonneau. 1998. Interaction of discourse structure with explicitness of discourse anaphoric noun phrases. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 327–358. Clarendon Press: Oxford.

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

J. Reynar. 1994. An automatic method of finding topic boundaaries. In *Proc. of 32nd ACL (Studen Session)*, pages 331–333.

M. Schreiner, B. Rehder, T. Landauer, and D. Laham. 1997. How latent semantic analysis (lsa) represents essay semantic content: Technical issues and analysis. In M. Shafto and P. Langley, editors, *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, page 1041. Mawhwah, NJ: Erlbaum.

C. Sidner. 1979. Toward a computational theory of definite anaphora comprehension in English. Technical Report No. AI-TR-537, Cambridge, Mass. MIT Press.

U. Turan. 1995. *Null vs. Overt Subjects in Turkish Discourse: A Centering Analysis*. Ph.D. thesis, University of Pennsylvania.

M. Walker. 1998. Centering: Anaphora resolution and discourse structure. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 401–35. Clarendon Press: Oxford.

B. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *NaturalLanguage and Cognitive Processes*, 6(2):107–135.

G. Youmans. 1991. A new tool for discourse analyis: The vocabulary-management profile. *Language*, 67:763–789.

HUM	E-R	ROUGH	PrH/E	PrH/E+R
6	5	15	5.074699	5.285312
6	6	22	5.992433	5.993116
6	6	15	5.992433	6.090629
6	6	22	5.992433	5.993116
6	6	24	5.992433	5.965255
6	4	22	4.156965	4.382483
6	4	13	4.156965	4.507857
6	6	28	5.992433	5.909533
6	5	30	5.074699	5.076356
6	4	30	4.156965	4.271039
6	4	0	4.156965	4.688952
6	5	20	5.074699	5.21566
6	6	21	5.992433	6.007046
6	6	50	5.992433	5.603064
6	6	25	5.992433	5.951324
6	5	21	5.074699	5.20173
6	6	6	5.992433	6.216003
6	5	35	5.074699	5.006704
6	5	25	5.074699	5.146008
6	5	30	5.074699	5.076356
5	4	15	4.156965	4.479996
5	5	7	5.074699	5.396756
5	4	5	4.156965	4.6193
5	5	38	5.074699	4.964912
5	4	40	4.156965	4.131735
5	5	45	5.074699	4.867399
5	6	27	5.992433	5.923464
5	4	30	4.156965	4.271039
5	5	21	5.074699	5.20173
5	5	16	5.074699	5.271382
5	5	20	5.074699	5.21566
5	6	32	5.992433	5.853811
5	4	40	4.156965	4.131735
5	4	10	4.156965	4.549648
5	4	23	4.156965	4.368552
5	5	20	5.074699	5.21566
5	6	25	5.992433	5.951324
5	4	25	4.156965	4.340691
5	5	50	5.074699	4.797747
5	6	10	5.992433	6.160281
4	3	11	3.239232	3.730401
4	5	45	5.074699	4.867399
4	4	46	4.156965	4.048152
4	3	50	3.239232	3.187114
4	3	36	3.239232	3.38214
4	3	33	3.239232	3.423931
4	5	42	5.074699	4.909191
4	3	50	3.239232	3.187114
4	4	36	4.156965	4.187457
4	4	40	4.156965	4.131735

HUM	E-R	ROUGH	PrH/E	PrH/E+R
4	3	11	3.239232	3.730401
4	3	75	3.239232	2.838853
4	4	38	4.156965	4.159596
4	3	62	3.239232	3.019949
4	4	12	4.156965	4.521787
4	4	40	4.156965	4.131735
4	5	48	5.074699	4.825608
4	3	9	3.239232	3.758262
4	3	81	3.239232	2.755271
4	3	100	3.239232	2.490593
3	3	55	3.239232	3.117462
3	4	30	4.156965	4.271039
3	4	81	4.156965	3.560587
3	4	42	4.156965	4.103874
3	3	50	3.239232	3.187114
3	3	66	3.239232	2.964227
3	3	42	3.239232	3.298558
3	2	40	2.321498	2.521102
3	3	75	3.239232	2.838853
3	3	40	3.239232	3.326418
3	3	78	3.239232	2.797062
3	3	62	3.239232	3.019949
3	2	55	2.321498	2.312145
3	2	30	2.321498	2.660406
3	3	?	3.239232	?
3	5	45	5.074699	4.867399
3	3	80	3.239232	2.769201
3	2	37	2.321498	2.562893
3	3	75	3.239232	2.838853
3	2	50	2.321498	2.381798
2	2	67	2.321498	2.14498
2	2	67	2.321498	2.14498
2	4	78	4.156965	3.602379
2	3	67	3.239232	2.950297
2	3	41	3.239232	3.312488
2	2	?	2.321498	?
2	1	67	1.403764	1.339664
2	2	20	2.321498	2.79971
2	2	42	2.321498	2.493241
2	2	50	2.321498	2.381798
1	2	50	2.321498	2.381798
1	2	0	2.321498	3.078319
1	1	67	1.403764	1.339664
1	3	71	3.239232	2.894575
1	3	57	3.239232	3.089601
1	0	100	0.48603	0.074643
1	1	85	1.403764	1.088916
1	1	67	1.403764	1.339664
1	2	57	2.321498	2.284285
1	1	0	1.403764	2.273002

Table 4: Table with the human scores (HUM), the *e-rater* scores (E-R), the Rough-Shift measure (ROUGH), the predicted values using *e-rater* as the only variable (PrH/E) and the predicted values using the *e-rater* and the added variable Rough-Shift (PrH/E+R). The ROUGH measure is the percentage of Rough-Shifts over the total number of identified transitions. The question mark appears where no transitions were identified.