Natural Language Engineering **1** (1): 000–000. Printed in the United Kingdom

Evaluation of Text Coherence for Electronic Essay Scoring Systems.

E. MILTSAKAKI
University of Pennsylvania, Philadelphia, PA, USA 19104

K. KUKICH †
Educational Testing Service, Princeton, NJ, USA 08541

(*Received 6 December 2002*)

Abstract

Existing software systems for automated essay scoring can provide NLP researchers with opportunities to test certain theoretical hypotheses, including some derived from Centering Theory. In this study we employ the Educational Testing Service's *e-rater* essay scoring system to examine whether local discourse coherence, as defined by a measure of Centering Theory's *Rough-Shift* transitions, might be a significant contributor to the evaluation of essays. *Rough-Shifts* within students' paragraphs often occur when topics are short-lived and unconnected, and are therefore indicative of poor topic development. We show that adding the *Rough-Shift* based metric to the system improves its performance significantly, better approximating human scores and providing the capability of valuable instructional feedback to the student. These results indicate that *Rough-Shifts* do indeed capture a source of incoherence, one that has not been closely examined in the Centering literature. They not only justify *Rough-Shifts* as a valid transition type, but they also support the original formulation of Centering as a measure of discourse continuity even in pronominal-free text. Finally, our study design, which used a combination of automated and manual NLP techniques, highlights specific areas of NLP research and development needed for engineering practical applications.

1 Introduction

The task of evaluating 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

† Current address for Karen Kukich: National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230

[1] http://lsa.colorado.edu.

[2] http://www.ets.org/research/erater.html

scores as frequently as two human scores match each other (Burstein, Kukich, Wolff, Chodorow, Braden-Harder, Harris and Lu 1998).

Essay scoring systems such as these can provide NLP researchers with opportunities to test certain theoretical hypotheses and to explore a variety of practical issues in computational linguistics. In this study, we employ the *e-rater* essay scoring system to test a hypothesis related to Centering Theory (Joshi and Weinstein 1981; Grosz 1983; Grosz, Joshi and Weinstein 1995). We focus on Centering Theory's *Rough-Shift* transition, which is the least well studied among the four transition types. In particular, we examine whether the discourse coherence found in an essay, as defined by a measure of relative proportion of *Rough-Shift* transitions, might be a significant contributor to the accuracy of computer-generated essay scores. Our positive finding validates the role of the *Rough-Shift* transition both in Centering Theory and in this application, and suggests a route for exploring Centering Theory's practical applicability to writing evaluation and instruction. Furthermore, our study design, which used a combination of automated and manual NLP techniques, highlights specific areas of NLP research and development needed for engineering practical applications.

Sections 2 and 3 of this paper briefly introduce the concepts of automated essay scoring and Centering Theory. Section 4 describes the central tenets of the Centering Model as employed in this study. Section 5 focuses specifically on the role of *Rough-Shift* transitions. Sections 6, 7, and 8 describe our *e-rater* Centering study and present its results followed by discussion. Finally, in Section 9 we discuss open issues and future work.

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, Kintsch and Landauer 1998; Schreiner, Rehder, Landauer and Laham 1997), employs a technique called Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer and Harshman 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 Educational Testing Service (ETS) writing evaluation experts for the essay portion of the Graduate Management Admissions Test(GMAT). The GMAT test is one of several criteria used by most U.S. graduate business schools to evaluate applicants. Over 200,000 GMAT tests are administered each year. Fully computerized, the GMAT test includes both a multiple choice section and an essay writing section. In the essay section, each examinee must compose two essays on general business-related topics randomly chosen by computer from a large pool of topics. Examinees are allowed 30 minutes to compose each essay and the average length of the essays is about 250 words. Essays are scored on a scale of 1 to 6 points, where a score of 1 indicates an extremely poor essay and a score of 6 indicates an excellent essay. Until recently, each essay was first scored by two trained writing evaluation experts. For those essays whose first two scores differ by more than one point (about ten percent), additional experts' scores are solicited. Starting in 1999, scores generated by *e-rater* were used in place of one of the first two experts, yielding a similar ten percent disagreement rate. The procedure for invoking additional experts as needed remains the same.

The essay scoring criteria used by GMAT writing evaluation experts, including, among others, syntactic variety, argument development, logical organization and clear transitions, are fully articulated in GMAT test preparation and scoring materials, which can be found at http://www.gmat.org. In the *e-rater* system, 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. Since Centering Theory was originally formulated as a measure of discourse coherence, we wondered whether it could be used to enhance *e-rater's* performance by adding a coherence feature to its evaluation criteria. To gain some initial insight, we first performed a preliminary study on a small sample of GMAT essays. We applied the Centering algorithm manually to a set of 32 essays, 8 from each of the top four levels, 6, 5, 4, and 3, counting the number of occurrences of each of the four types of Centering transitions (*Continue*, *Retain*, *Smooth-Shift* and *Rough-Shift*) in each essay. We observed that essays that received higher scores by writing experts tended to have significantly lower percentages of Centering Theory's *Rough-Shift* transitions than essays with lower scores. Specifically, for 15 of the 16 essays scored 5 or 6, less than 25 percent of the total number of transitions were *Rough-Shifts*, while the percentage of *Rough-Shift* transitions was greater than 40 percent for almost all of the essays scored 3 or 4. None of the other three Centering transition types showed either a positive or negative pattern across essay scores. A detailed account of the essay scores and transition counts can be found in Table (5) in the Appendix. This observation encouraged us to undertake a fuller study to explore the hypothesis that the Centering Model provides a reasonable measure of coherence (or lack of) reflecting the evaluation performed by writing experts. Specifically, in the study described here, we investigate the effect of adding a *Rough-Shift* percentage feature to *e-rater*'s existing array of features.

3 Overview of Centering

Two different lines of work were mainly responsible for the development of Centering Theory. Originally, Joshi, Kuhn, and Weinstein (Joshi and Kuhn 1979; Joshi and Weinstein 1981) proposed Centering as a model of the complexity of inferencing involved in discourse when speakers process the meaning of an utterance and integrate it into the meaning of the previous discourse. Grosz and Sidner (Sidner 1979; Grosz 1977; Grosz and Sidner 1986) recognized what they called the 'attentional state' as a basic component of discourse structure and proposed that it consisted of two levels of focusing: global and local. For Grosz and Sidner, Centering Theory provided a model for monitoring local focus. A synthesis of these two approaches yielded the Centering Model which was designed to account for those aspects of processing that are responsible for the difference in the perceived coherence of discourses as those demonstrated in (1) and (2) below (Grosz, Joshi and Weinstein 1995).

(1) a. John went to his favorite music store to buy a piano.
 b. He had frequented the store for many years.
 c. He was excited that he could finally buy a piano.
 d. He arrived just as the store was closing for the day.

(2) a. John went to his favorite music store to buy a piano.
 b. It was a store John had frequented for many years.
 c. He was excited that he could finally buy a piano.
 d. It was closing just as John arrived.

Discourse (1) is intuitively more coherent than discourse (2). This difference may be seen to arise from the different degrees of continuity in what the discourse is about. Discourse (1) centers a single individual (*John*) whereas discourse (2) seems to focus in and out on different entities (*John, store, John, store*). Centering is designed to capture these fluctuations in continuity.

4 The Centering Model

In this section, we present the basic definitions and assumptions in Centering as discussed in the literature (Walker, Joshi and Prince 1998, among others). We specify and motivate the ones that we have made in this study.

Discourse consists of a sequence of textual segments and each segment consists of a sequence of utterances. In Centering Theory, utterances are 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. (Ranking is described in Section 4.4.) 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 the previous utterance, $Cf(U_{i-1})$, that is *realized* in the current utterance, U_i, is its designated BACKWARD-LOOKING CENTER, Cb. The BACKWARD-LOOKING CENTER is a special member of the Cf set because it represents the discourse entity that U_i is about, what in the literature is often called the 'topic' (Reinhart 1981; Horn 1986).

The Cp for a given utterance may be identical with its Cb, but not necessarily so. Depending on the identity relations among Cb's and Cp's in subsequent utterances, four different types of transitions are defined, *Continues*, *Retains*, *Smooth-Shifts* and *Rough-Shifts*. The key element in computing local coherence in discourse is the distinction between looking back in the discourse with the Cb and projecting preferences for interpretation in the subsequent discourse with the Cp.

4.1 Discourse segments

Segment boundaries are extremely hard to identify in an accurate and principled way. Furthermore, existing segmentation 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. The same is true for work done in the Centering framework. Passonneau (1998), for example, implemented Centering to detect segment boundaries. The rationale of her approach was that assuming coherent texts, *Rough-Shifts* can be used to locate segment boundaries. 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. The *Rough-Shift* based metric that we propose evaluates textual coherence *within* each paragraph in an essay. The final score is summative, adding up the coherence evaluation of each paragraph. In other words, we compute the degree of coherence within each segment and then we compute a single score for all the segments in an essay. Our metric does not compute textual coherence *across* segments.

4.2 Centering transitions

Four types of transitions, reflecting four degrees of coherence, are defined in Centering. They are *Continue*, *Retain*, *Smooth-Shift*, and *Rough-Shift*. Their order of precedence is shown in transition ordering rule (1). The rules for computing the transitions are shown in Table (1). For a segment initial U_{i-1} with Cb=none, we assume $Cb(U_i)=Cb(U_{i-1})$ for the computation of the Centering transition in U_i.

Table 1. *Table of transitions*

	$Cb(U_i)=Cb(U_{i-1})$	$Cb(U_i)\neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	*Continue*	*Smooth-Shift*
$Cb(U_i)\neq Cp(U_i)$	*Retain*	*Rough-Shift*

For a segment medial U_{i-1} with Cb=none, we assume $Cb(U_i)\neq Cb(U_{i-1})$ for the computation of the Centering transition in U_i.

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

Centering defines one more rule, the Pronoun rule, which we will discuss in detail in Section 5.

4.3 Utterances

In Centering Theory, a single Centering update unit is the 'utterance', but in early formulations of Centering Theory, the 'utterance' was not defined explicitly. Some researchers have defined the utterance as a single tensed clause, but in this work we take the utterance to be the main clause and all its associated tensed dependent clauses. We discuss the rationale for this assumption here.

Kameyama (1998) defined the 'utterance' as, roughly, the tensed clause, with relative clauses and clausal complements as exceptions. Kameyama's motivation for treating adverbial subordinate clauses as independent units came from backward anaphora. She argued that treating adverbial subordinate clauses as independent units predicts that a pronoun in a fronted subordinate clause, as in (3c) for example, is anaphorically dependent on an entity already introduced in the immediate discourse and not on the subject of the main clause it is attached to:

(3)
a. Kern_i began reading a lot about the history and philosophy of Communism
b. but never 0_i felt there was anything he as an individual could do about it.
c. When he_i attended the Christina Anti Communist Crusade school here about six months ago
d. Jim_i became convinced that an individual can do something constructive in the ideological battle
e. and 0_i set out to do it.

This view on backward anaphora, in fact, was strongly professed by Kuno (1972), who asserted that there was no *genuine* backward anaphora: the referent of an apparent cataphoric pronoun must appear in the previous discourse. Kameyama's argument (also Kuno's) is weak in two respects. First it is not empirically tested that in cases of backward anaphora the antecedent is found in the immediate discourse.

Carden (1982) and van Hoek (1997) provide empirical evidence of pronouns which are the first mention of their referent in discourse. Most recently, Tanaka (2000) reported that in the cataphora data retrieved from the Anaphoric Treebank, out of 133 total occurrences of personal pronouns encoded as 'cataphoric', 47 (35.3%) were 'first mentioned'. Among the 47 cases of 'first mention' cataphora, 6 instances were discourse initial.[3]

Based on cross-linguistic corpus studies (Miltsakaki 1999) and experimental work in English and Greek (Miltsakaki 2001; Miltsakaki 2002a), Miltsakaki (2002b), argues that for the purposes of discourse organization and textual coherence, subordinate clauses are not processed as independent processing units. She argues that treating subordinate clauses as independent units ('utterances') yields counterintuitive topic transitions. This can be seen for English in the constructed example shown in (4).

(4) Sequence: Main-subordinate-main

a. John had a terrible headache.
   Cb= ?, Cf= John>headache, Transition=none
b. When the meeting was over,
   Cb=none, Cf= meeting, Transition=*Rough-Shift*
c. he rushed to the pharmacy store.
   Cb=none, Cf=John, Transition=*Rough-Shift*

Allowing the subordinate clause to function as a single update unit yields a sequence of two *Rough-Shifts*, which is diagnostic of a highly discontinuous discourse. Further, if indeed there are two *Rough-Shift* transitions in this discourse the use of the pronominal in the third unit is puzzling. A sequence of two *Rough-Shift* transitions in this short discourse is counterintuitive and unexpected given that of all Centering transitions, *Rough-Shifts* in particular have been shown to a) disfavor pronominal reference (Walker, Iida and Cote 1994; Di Eugenio 1998; Miltsakaki 1999, among others), and b) be rare in corpora, to the extent that the transition has been ignored by some researchers (Di Eugenio 1998; Hurewitz 1998, among others).

In addition, simply reversing the order of the clauses, as shown in (5), causes an unexpected improvement with one *Rough-Shift* transition being replaced with a *Continue*.

(5) Sequence: Main-main-subordinate

a. John had a terrible headache.
   Cb=?, Cf=John>headache, Transition=none

[3] The Anaphoric Treebank is a corpus of news reports, annotated, among other things, with type of anaphoric relations. The Anaphoric Treebank is developed by UCREL (Unit for Computer Research on the English Language) in Lancaster University, collaborating with IBM T.J. Watson Research Center, Yorktown Heights, New York.

b. He rushed to the pharmacy store
Cb=John, Cf=John>pharmacy store, Transition=*Continue*

c. when the meeting was over.
Cb=none, Cf=meeting, Transition=*Rough-Shift*

Assuming that the two discourses demonstrate a similar degree of continuity in the topic structure (they are both *about* 'John'), we would expect the transitions to reflect this similarity when, in fact, they do not. Presumably, the introduction of a new discourse entity, 'meeting', in the time-clause does not interfere with discourse continuity, nor does it project a preference for a shift of topic, as the Cp normally does when it instantiates an entity different from the current Cb.

According to Miltsakaki (2002b), the 'utterance', a single Centering update unit, consists of one main clause and all its associated tensed dependent clauses, including sentential complements of verbs, relative clauses and subordinate clauses. In Miltsakaki (2002b), the term 'subordinate clause' is used to describe tensed, adverbial clauses introduced by a subordinate conjunction (e.g., when, because, as soon as, although etc.). To identify subordinate clauses the 'reversibility test' is applied (Quirk, Greenbaum, Leech and Svartvik 1972); subordinate clauses can be preposed with respect to the main clause. For example, in (6), *although* is classified as a subordinate conjunction and the although-clause is classified as a subordinate clause because placing the although-clause before the main clause retains grammaticality. Conversely, *however* in (8) is not classified as a subordinate conjunction because preposing the clause it is associated with yields ungrammaticality.

(6) John traveled by air although he is afraid of flying.

(7) Although he is afraid of flying, John traveled by air.

(8) John traveled by air. However, he is afraid of flying.

(9) # However, he is afraid of flying. John traveled by air.

Returning to our mini discourses in (4) and (5), if we process the subordinate clause in the same unit as the main clause, but allow main clause entities to rank higher than subordinate clause entities, we compute a *Continue* transition independent of the linear position of the subordinate clause (see further discussion about the Cf ranking in Section 4.4). The computation is shown in (10).

(10) a. John had a terrible headache.
Cb=?, Cf=John>headache, Transition=none

b. When the meeting was over, he rushed to the pharmacy store.
Cb=John, Cf=John>pharmacy store>meeting, Transition=*Continue*

Interestingly, defining the 'utterance' as the unit consisting of a main clause and its dependent subordinate clauses is further supported by the 'SX because SY' type of complex sentence studied by Suri, McCoy and DeCristofaro (1999). Suri et al (1999) developed an anaphora resolution algorithm (RAFT/RAPR) based on models of discourse salience. In their algorithm, they introduce the 'Prefer SX' hypothesis to account for the fact that in (13) the referent of *he* is *the ex-convict* and not *Dodge*, which is the most recent subject, i.e., the subject of the because-clause. If (S2) is treated as one unit, then the referent of *he* will correctly resolve to *the ex-convict* which would be the highest ranked entity of the preceding unit.

(11) (S1) Dodge was robbed by an ex-convict the other night.

(12) (S2) The ex-convict tied him up because he wasn't cooperating.

(13) (S3) Then he took all the money and ran.

In what follows, we adopt the definition of the 'utterance', the single Centering update unit, suggested in Miltsakaki (2002b): *a single Centering update unit consists of a main clause and all its associated dependent clauses.* Following Kameyama (1998), non-tensed clauses are assumed to belong to the clause containing them.

4.4 Cf ranking.

As mentioned earlier, the PREFERRED CENTER of an utterance is defined as the highest ranked member of the Cf set. The ranking of the Cf members is determined by the salience status of the entities in the utterance and may vary across language. Kameyama (1985) and Brennan, Walker-Friedman and Pollard (1987) proposed that the Cf ranking for English is determined by grammatical function as follows:

(2) Rule for ranking of forward-looking centers:

SUBJ>IND. OBJ>OBJ>OTHERS

Later crosslinguistic studies based on empirical work (Di Eugenio 1998; Turan 1995; Kameyama 1985) determined the following detailed ranking, with QIS standing for quantified indefinite subjects (people, everyone, etc.) and PRO-ARB (we, you) for impersonal pronominals.

(3) Revised rule for the ranking of forward-looking centers:

SUBJ>IND. OBJ>OBJ>OTHERS>QIS, PRO-ARB.

We assumed the Cf ranking given in (3). As suggested by a reviewer, the content and the ranking of the Cf list may also vary across different types of essays within the same language. Indeed we have made a few modifications to reflect the properties of the type of essay under investigation. We will turn to those shortly. Overall, though, the Cf ranking in (3) worked well for the GMAT essays. This is because text coherence in students' paragraphs was often achieved by centering a certain individual or concept as shown in (14).

(14) Another example of an individual who has achieved success in the business world through the use of conventional methods is **Oprah Winfrey**. One may not think of her as a 'businesswoman', however **she** has managed to install her own production company, all done through hard work and perseverance. Indeed, perseverance is a time honoured method of gaining success. **She** has indeed been able to persevere through all the obstacles which she had to face throughout her career. It is because of this hard work and perseverance (again, conventional practices), that **she** has been able to attain her success.

To construct the ranking of the Cf list under the assumption that the 'utterance' contains both a main clause and its subordinate clauses, we assume the augmented Cf **ranking rule** proposed in Miltsakaki (2002b). The 'M' prefix stands for main clause and the 'S_n' prefix stands for the nth subordinate clause. The relevant ranking of the various types of subordinate clauses is currently left unspecified. In our study, the relevant ranking of subordinate clauses was never crucial. In our study, the 'S' in the adopted *augmented ranking rule*, stands for any tensed dependent clause.

Augmented ranking rule
M-Subject > M-indirect object > M-direct object > M-other > M-QIS,Pro-ARB > S1-subject > S1-indirect object > S1-direct object > S1-other > S1-QIS,Pro-ARB > S2-subject > ...

Notice that the **augmented ranking rule** is insensitive to the linear order of the subordinate clauses. While no corpus study has yet been conducted to test whether the insensitivity of the rule to linear order is justified, there is accumulating evidence pointing to this direction across languages (Miltsakaki 2002b). First, the augmented ranking rule points to interesting new directions in understanding backward anaphora. With the exception of a few modal contexts shown in (17),[4] backward anaphora is most commonly found in preposed subordinate clauses, (15), and not in sequences of main clauses, (16). From the proposed 'utterance' definition and the augmented ranking rule it follows that surface backward anaphora is no longer 'backward' once the Cf list is constructed and ranked. The referent of the pronoun in such cases appears lower in the Cf list ranking and, in fact, looks backwards for an antecedent as any other normal pronoun would. To illustrate the point, the Cf list for (15) contains *John>shower>he-referent*. The pronoun looks back for an antecedent, intra-sententially, and resolves to the only compatible antecedent available, *John*.

(15) As soon as he_i arrived, $John_i$ jumped into the shower.

(16) #He_i arrived and $John_i$ jumped into the shower.

(17) He_i couldn't have imagined it at the time but $John_i$ turned out to be elected President in less than 3 years.

Further evidence comes from cross-linguistic observations on anaphora resolution. The most striking examples come from Japanese.[5] In Japanese, topics and subjects are lexically marked (wa and ga respectively) and null subjects are allowed. Note that subordinate clauses must precede the main clause. Consider the Japanese discourse (18). Crucially, the referent of the null subject in the second main clause resolves to the topic marked subject of the first main clause, ignoring the subject-marked subject of the intermediate subordinate clause.

[4] Thanks to Ellen Prince for pointing out this example.

[5] Thanks to Kimiko Nakanishi for providing these data. In a Centering study she conducted in Japanese she also concluded that treating subordinate clauses as independent units would yield a highly incoherent Japanese discourse.

(18) a. Taroo wa tyotto okotteiru youdesu
Taroo TOP a-little upset look
'Taroo looks a little upset.'

b. Jiroo ga rippana osiro o tukutteiru node
Jiroo SUB great castle OBJ is-making because
'Since Jiroo is making a great castle,'

c. ZERO urayamasiino desu
ZERO jealous is
'(He-Taroo) is jealous.'

We saw a similar case in English in the previous section, example (4). Similar cases have also been identified in a small corpus study in Greek (Miltsakaki 2001a). In light of these observations, we made the decision to adopt the augmented ranking rule as our working hypothesis. We are currently in the process of conducting experimental and corpus-based studies to empirically evaluate the hypothesis for English and other languages.

Returning to the Cf ranking in our study, a modification we made involved the status of the pronominal *I*. [6] We observed that in low-scored essays the first person pronominal *I* was used extensively, normally presenting personal narratives. Such extensive use of *I* in the subject position produced an effect of high coherence. However, because personal narratives were unsuited to this essay writing task, they were assigned lower scores by expert readers. 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.

A further modification we made to the Cf ranking involved constructions containing the verb *to be*. In these constructions (e.g., Another company would be Gerber..., There is more promise ...), we ranked the noun phrase following the verb *to be* higher than its structural subject. Our rationale for this modification is as follows.

The verb *to be* appears in two types of constructions: specificational and predicational. The modification is relevant only for the specificational cases. The predicational *be* in, for example, the sentence *John is happy/a doctor/the President of the United States*, does not make any semantic contribution. The post verbal nominal phrase forms the predicate of the sentence and assigns a property holding of *John*. It does not introduce another entity distinct from *John*.

The specificational *be*, as in *The cause of his illness is this virus here*, is a predicate of identity or equation (Heycock and Kroch 1997). It is in these cases that

[6] In fact, a similar modification has been 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.

we rank the post verbal nominal higher than the subject. In (19), for example, *Oprah Winfrey* is the highest ranked entity in the Cf list because the verb *to be* is specificational.

(19) Another example of an individual who has achieved success in the business world through the use of conventional methods is Oprah Winfrey.

Finally, expletives do not evoke discourse entities and therefore do not participate in the Cf list. In (20), for example, the highest ranked entity is *success*.[7]

(20) It is possible to achieve real success in business by following conventional methods.

4.4.1 Complex NPs

In the case of complex NPs, which have the property of evoking multiple discourse entities (e.g. his mother, software industry), the working hypothesis commonly assumed (e.g., (Walker and Prince 1995)) is ordering from left to right.[8] With respect to complex NPs containing *possession* relationships the following clarification is in order. English has two types of *possessive* constructions. The first construction is the *genitive* construction realized with an apostrophe plus the letter *s* at the end of the noun. In this construction, the *possessor* is to the left of the *possessee*, for example *Mary's father*. The second construction contains the preposition *of*. In this case, the *possessor* is to the right of the *possessee*. To maintain uniformity for the ranking of the complex NP, we assume linearization of the complex NP according to the genitive construction and then rank from left to right. In (21b), for example, *TLP* ranks higher than both *success* and the *secret*. The ranking is easy to see if we linearize *The secret of TLP's success* to *TLP's success's secret*.

(21) a. Trade & Leisure Publications is a successful publishing house in Russia, with two market-leading monthly consumer magazines.

b. The secret of TLP's success, however, is not based on developing or exploiting some new technology or business strategy.

c. Rather, TLP follows a business strategy that has been known since business began.

5 The role of *Rough-Shift* transitions

Although Centering Theory was originally formulated as a model of discourse organization, to date most research has focused on its applicability to the tenacious problem of pronoun resolution. As mentioned briefly earlier, the Centering Model includes one more rule, the Pronoun Rule given in (4).

[7] In accordance with the Cf ranking rule (3), the subject of the infinitival construction *to achieve* is ranked low because it is a non-referential indefinite noun phrase.
[8] But see also Di Eugenio (1998) for the treatment of complex NPs in Italian.

(4) Pronoun Rule: If some element of $\mathrm{Cf}(\mathrm{U}_{i-1})$ is realized as a pronoun in U_i, then so is the $\mathrm{Cb}(\mathrm{U}_i)$.

The Pronoun Rule reflects the intuition that pronominals are felicitously used to refer to discourse-salient entities. As a result, Cbs are often pronominalized, or even deleted (if the grammar allows it). Rule (4) then predicts that if there is only one pronoun in an utterance, this pronoun must realize the Cb. The Pronoun Rule and the distribution of forms (definite/indefinite NPs and pronominals) over transition types plays a significant role in the development of anaphora resolution algorithms in NLP.

Note that the utility of the Pronoun Rule and the Centering transitions in anaphora resolution algorithms relies heavily on the assumption that the texts under consideration are maximally coherent. In maximally coherent texts, however, *Rough-Shift* transitions are rare, and even in less than maximally coherent texts they occur infrequently. For this reason the distinction between *Smooth-Shifts* and *Rough-Shifts* was collapsed in previous work (Di Eugenio 1998; Hurewitz 1998). The status of *Rough-Shift* transitions in the Centering Model was therefore unclear, receiving only negative evidence: *Rough-Shifts* are valid because they are found to be rare in coherent discourse.

In this study we gain insights pertaining to the nature of the *Rough-Shifts* precisely because we are forced to drop the coherence assumption. After we applied the Centering algorithm and computed a *Rough-Shift* coherence measure for 100 student essays as described in detail in the next section, we observed a crucial insight. Namely, in our data, the incoherence detected by the *Rough-Shift* measure is *not* due to violations of Centering's Pronominal Rule or infelicitous use of pronominal forms in general.

Table (2) shows the distribution of nominal forms over *Rough-Shift* transitions. Out of the 211 *Rough-Shift* transitions found in the set of 100 essays, in 195 instances, the preferred center (or Cp as indicated in the rules in Table 1) was a nominal phrase, either definite or indefinite.

Pronominals occurred in only 16 instances, of which 6 cases instantiated the pronominals 'we' or 'you' in their impersonal sense. These data strongly indicate that the incoherence found in student essays is not due to the processing load imposed on the reader to resolve anaphoric references. Instead, the incoherence in the essays is due to discontinuities caused by introducing too many undeveloped topics within what should be a conceptually uniform segment, i.e., the paragraph. This is, in fact, what the *Rough-Shift* measure picked up. In the next section we show that *Rough-Shift* transitions provide a reliable measure of *incoherence*, correlating well with scores provided by writing experts.

These results not only justify *Rough-Shifts* as a valid transition type but they also support the original formulation of Centering as a measure of discourse continuity even when anaphora resolution is not an issue. It seems that *Rough-Shifts* are capturing a source of incoherence that has been overlooked in the Centering literature. The processing load in the *Rough-Shift* cases reported here is not increased by the effort required to resolve anaphoric reference but instead by the effort required to find the relevant topic connections in a discourse bombarded with a rapid

Table 2. *Distribution of nominal forms over* Rough-Shifts

	Def. Phr.	Indef. Phr.	Prons	Total
Rough-Shifts	75	120	16	211
Total	195		16	211

succession of multiple entities. That is, *Rough-Shifts* are the result of absent and extremely short-lived Cbs. We interpret the *Rough-Shift* transitions in this context as a reflection of the incoherence perceived by the reader when s/he is unable to identify the topic (focus) structure of the discourse. This is a significant insight which opens up new avenues for practical applications of the Centering Model.

6 The *e-rater* Centering study

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*. Our data consist of student essays whose degree of coherence is under evaluation and therefore cannot be assumed.

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 (the same set of 100 essays also mentioned in Section 5) covering the full range of the scoring scale, where 1 is lowest and 6 is highest, as shown in Tables (6) and (7) in the Appendix. Using students' paragraph marking as segment boundaries (for reasons specified in Section 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 the *e-rater*'s performance.[9] Although the *Rough-Shift* measure itself is simple, its automatic computation raises some interesting research challenges which are discussed here.

6.1 Implementation

For this study, we decided to manually tag coreferring expressions despite the availability of coreference software. We made this decision because a poor performance of the coreference software would give us distorted results and we would not be able to test our hypothesis. Similarly, we manually tagged Preferred Centers (as Cp's) for the same reason. We are aware of the difficulties that can arise with regard to manual annotation and inter-annotator agreement, and we address this issue in the next section. We also manually tagged other entities in utterances, but we only needed to mark them as OTHER, since this information is sufficient for the

[9] Tables (8), (9), and (10) in the Appendix show the counts of the Centering transitions computed for each of the 100 essays.

automatic computation of the Cb and all of the transitions indicated in Table (1). From a natural language engineering perspective, this work highlights the need for more research and development toward reliable named-entity recognizers, coreference resolvers, and software needed to determine Cf ranking, for example syntactic parsers and semantic role identifiers.

Discourse segmentation and the implementation of the Centering algorithm for the computation of the transitions were automated. Segment boundaries were automatically marked at paragraph breaks, and transitions were computed according to the rules given in Table (1). As output, our system computed 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.

6.2 Inter-annotator agreement

Manually annotating corpora for specific linguistic features is known to be fraught with difficulties. See Poesio and Vieira (1998) for an excellent account of the issues regarding annotating for definite descriptions. As mentioned in the previous section, we chose to manually annotate essays to identify co-referring expressions and Cp's because truly robust and accurate software for these tasks does not yet exist. Indeed, this is an active and important area of research. We believed that manual tagging by the authors would produce more reliable data, especially since the Cp is a well-defined concept and we did not expect high disagreement. As a reality check for this belief, we performed a small inter-annotator agreement study. We randomly extracted five essays from each of the six scoring levels in our study set of 100 essays. We used this set of thirty essays to compare inter-annotator agreement. Each author independently tagged only the Cp in each utterance of these thirty essays in accordance with the Cf ranking rule given in Section 4.4. The thirty essays of this inter-annotation set contained 444 utterances.

For the total of 444 annotated Cps, the two annotators were in agreement in 405 cases, that is in 91% of all utterances. In 39 cases the two annotators marked a different noun phrase as the Cp. To examine the effect of the Cp mismatch, we looked at those cases to check if the transition change involved *Rough-Shifts*. For 31 of the 39 cases of Cp mismatch, choosing a different Cp did not affect the computation of the transition. This is because in most of these cases no Cb was identified in the subsequent utterance, so the Cp of the current utterance did not matter. For 7 of the 8 cases where the Cp mismatch would change the transition, the change involved *Continue*, *Retain* and *Smooth-Shift* transitions (for example, changing a *Continue* to a *Retain* or *Smooth-Shift* and so on). In only one case would the transition change from a *Smooth-Shift* to a *Rough-Shift*, thus affecting the value of the *Rough-Shift* metric for that essay. The results of the inter-annotator study and the close inspection of the effect of the mismatches were very encouraging. In effect, only one case out of the 444 would affect the value of the *Rough-Shift* metric. To further validate our use of manual tagging, we computed the Kappa statistic for our small study. In the following section, we discuss our Kappa statistic computation.

6.2.1 The Kappa statistic

The Kappa statistic (Cohen 1960; Kraemer 1982), introduced to NLP by Carletta (1996) for corpus annotation, has been widely used in the field as a measure of inter-annotator agreement. The Kappa calculation provides a statistical method to correct for chance agreement among annotators. For Kappa > 0.8 annotation is considered reliable. For $Kappa < 0.68$, annotation is considered unreliable. Values in between may allow some tentative conclusions to be drawn (Poesio and Vieira 1998).

The usefulness of the Kappa statistic to quantify levels of agreement has been questioned, however (Maclure and Willett 1988; Guggenmoos-Holzmann 1993). The criticism is that the Kappa computation is reliable only in cases where the statistical independence of raters is guaranteed, and raters are by definition dependent because they all rate the same cases according to a pre-specified rule. Critics point out that 'Lacking an explicit model of decision-making, it is not clear how chance affects the decisions of actual raters and how one might correct for it.'[10] Keeping these concerns in mind, we find it useful to compute the Kappa statistic as a the means to compare with Kappa statistics that have been reported in other inter-annotator studies.

The formula for the computation of Kappa is given below:

$$K = \frac{P(A) - P(E)}{1 - P(E)},$$

where P(A) is the proportion of times the annotators agree and P(E) is the proportion of times that we would expect the annotators to agree by chance.[11] To compute the P(E), Poesio and Vieira (1998) give the formula:

$$P(E) = (\frac{\text{number of instances of classification category}}{\text{total number of classification judgments}})^2.$$

To compute P(E) in our case, we observed that the probability of an annotator correctly tagging the Cp is the probability of picking the correct NP out of all the NPs in an utterance. So we computed the average number of NP's for each utterance (by dividing the total number of NP's by the total number of utterances). The average number of NP's per utterance is 4.83. The chance probability of two annotators tagging the same NP as the Cp is $(1/4.83)^2$. P(A) is the percentage agreement for all descriptions, 0.91 in our case. The final computation is given below.

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.91 - 0.04)}{(1 - 0.04)} = \frac{0.87}{0.96} = 0.91.$$

A Kappa of .91 indicates very good inter-annotator reliability, as we expected for this relatively simple task.

[10] http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm and references therein.

[11] For the details of the formula, its description and its computation we have consulted (and replicated) the excellent presentation of the Kappa statistic in Poesio and Vieira (1998).

This simple study was perhaps even more useful in that it helped us identify causes of disagreement that can be used to further refine a future algorithm for the identification of a Cp. We found that our disagreement instances fell in two main groups. The first group contained instances where there was some apparent confusion as to the ranking of phrases such as *a person, people, impersonal 'we' and 'they'*, etc. with respect to other indefinite phrases. For example, in (23), one annotator picked *they* as the referent because it was the subject of the sentence. The other picked *rich or lasting success* because *they* referred to *the person*, which is impersonal.

(22) However, real success can be measured depending on what the person wants out of life.

(23) How they define rich or lasting success.

The second group contained cases with *I* as one of the potential Cps. Apparently, it was unclear whether all *I's* were to be ignored, or just the *I's* in the constructions *I think, I believe, I agree*, etc. For example, in (24), one annotator picked *I* as the Cp and the other picked *the service*.

(24) I do not do so because the service has unconventional way of couriering documents.

6.3 An example of coherent *text*

What follows is a small excerpt (a paragraph) of a student essay scored 6.[12][13] For each utterance, enclosed in the <UT-n> 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-1> 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-2><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-3>If they opted for an extreme market-place, <CP REF='3'>they</CP> would be ousted.</UT> Cb=3 Cp=3 Tr=*Continue*
<UT-4>Just look who <CP REF='3'>their</CP> market is!</UT> Cb=3 Cp=3

[12] Only proper names have been changed for privacy protection. Spelling and other typographical errors have been corrected, also for privacy reasons.
[13] In this and the following example, the identified transitions evaluate the degree of (in)coherence in the quoted paragraphs. This evaluation may not reflect the final score of the essay. The final (in)coherence score for the essay as a whole is based on the sum of the scores of all the paragraphs contained in that essay.

Tr=*Continue*
<UT-5>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-6><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-7><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*

In the first utterance, the *Famous name's Baby Food* is marked as the Cp because it appears in a main clause, after the verb *to be* in a specificational construction (see Section 4.4). In the second utterance, *this company* is marked as the Cp because it is the subject of the main clause. Similarly, in the third utterance, the referent of *they* is the Cp because it is the subject of the main clause. In the fourth utterance, the implicit subject of imperative form, the impersonal *you*, is ignored, so the referent of *their* is the Cp because it is the highest ranked entity in the complex NP *their market*, following the rule for ranking entities in complex NPs from left to right as explained in Section 4.4.1. In the fifth utterance, the first entity in the complex NP in the subject role, *the Famous name*, is the Cp following the left-to-right ranking of entities in complex NPs. In the sixth and seventh utterances, *Famous name* is the Cp because in both cases it realizes the subject of the main clause.

6.4 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 captures the fact that the introduced centers are poorly developed.

<UT-8>I disagree with <CP REF='1'>the opinion</CP> stated above.</UT> Cb=none Cp=1 Tr=none
<UT-9>In order to achieve <CP REF='4'>real and lasting success</CP> <OTHER REF='2'>a person</OTHER> does not have to be a billionaire.</UT> Cb=none Cp=4 Tr=*Rough-Shift*
<UT-10>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*

In utterance 8, the referent of *I* is ignored and the only other entity realized in the utterance is marked as the Cp. In utterance 9, there is only a main clause, as the infinitive *in order to achieve* is not a tensed clause and therefore does not count as a separate subordinate clause according to our definition. The subject of the main clause, *a person*, ranks lower than the other entities in the utterance because it is an indefinite, non specific, non-referential NP. Furthermore, the verb *to be* in the main clause is predicational and therefore the NP *a billionaire* does not evoke an entity. The subject of the infinitive, the impersonal *you*, is not retrieved. The remaining NP *real and lasting success* is marked as the Cp. In utterance 10, the only available subject *conventional practices* is marked as the Cp.

7 Study results

A summary of the results of applying the Centering algorithm to 100 GMAT essays is shown in Table (3). The first column in Table (3), labeled HUM, indicates the score level of the essays as graded by human raters. The second column, labeled E-R, gives the average *e-rater* score for all essays at each (human) score level. There were twenty essays each for score levels 6, 5, 4, and 3, and ten essays each for score levels 2 and 1, totaling 100 essays. The third column, labeled ROUGH, shows the average *Rough-Shift* measure at each score level. The full details of the human scores, *e-rater* scores and *Rough-Shift* measure for each of the 100 essays are shown in Table (6) and Table (7) in the Appendix.

HUM	E-R	ROUGH
6	5.25	22.7
5	4.8	24.95
4	3.6	43.25
3	3	54.37
2	2.33	55.44
1	1.6	55.40

Table 3. *Summary table with average E-R and ROUGH scores for each essay score.*

Comparing columns HUM and ROUGH in Table (3), we observe that essays with scores from the higher end of the scale tend to have lower percentages of *Rough-Shifts* than those from the lower end, repeating the same pattern we observed in our preliminary study of 32 essays. To statistically evaluate whether this observation can be used to improve *e-rater*'s performance, we regressed the variable X=ROUGH (the predictor) by Y=HUM. As expected, the regression yielded a negative coefficient (ROUGH=0.013) for the ROUGH predictor, thus penalizing occurrences of *Rough-Shifts* in the essays. It also yielded a highly significant p-value ($p<0.0013$) on the t-test for ROUGH for these 100 essays, suggesting that adding the variable ROUGH to the *e-rater* model can contribute to the accuracy of the model. [14] The

[14] The t ratio is formed by first finding the difference between the estimate and the hypothesized value and then dividing that quantity by its standard error. A significant t

magnitude of the contribution indicated by this regression is approximately 0.5 point, a reasonably sizable effect given the scoring scale (1-6).

Additional work is needed to precisely quantify the contribution of ROUGH. Ideally, we would incorporate the variable ROUGH into the building of a new *e-rater* scoring model and compare the results of the new model to the original *e-rater* model. Because we could not modify the original *e-rater* model directly, we used a standard statistical technique known as jackknifing (Becker and Chambers 1984; Mosteller and Tukey 1977) to simulate the effect of incorporating the ROUGH variable into an *e-rater* model. Jackknifing calls for repeatedly using a random portion of a data set to predict values for the unused portion and averaging over all subset predictions to estimate a whole set prediction. We performed 100 tests with ERATER as the sole variable, leaving out one essay each time, and recorded the prediction of the model for that essay. Then we repeated the procedure using both the ERATER and ROUGH variables. This procedure enabled us to estimate the scores predicted by both *e-rater* alone and *e-rater* enhanced with a *Rough-Shift* measure.

The predicted values for ERATER alone and ERATER+ROUGH are shown in columns E(PRED) and E+R(PRED)in Table (4).

HUM	E(PRED)	E+R(PRED)
6	5.29	5.36
5	4.89	4.98
4	3.78	3.75
3	3.24	3.12
2	2.63	2.59
1	1.97	2.03

Table 4. *Summary table with E(PRED) and E+R(PRED) scores for each essay score level.*

As can be seen by comparing the columns E(PRED) and E+R(PRED), the addition of the *Rough-Shift* measure moved the e-rater score closer to the human score for levels 6, 5, 4, and 2.

By examining the detailed comparisons of predictions for each of the 100 essays, shown in Table (6) and Table (7) in the Appendix, we observe that, indeed, 57 % of the predicted values shown in the E+R(PRED) column are better approximations of the human scores, especially in the cases where the *e-rater* score differs by 2 or more points from the human score. In all these cases, the E+R(PRED) value unmistakably tilts the predicted score in the right direction. In summary, the results clearly indicate a greater agreement with human expert scores using a *Rough-Shift* enhanced version of *e-rater*.

ratio indicates that for the tested variable the null hypothesis must be rejected. In our case, the t ratio indicates that the ROUGH variable is significant.

8 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. Note, however, that overall scores and coherence scores need not be strongly correlated. Indeed, our data contain several examples of essays with high coherence scores, i.e., low percentages of *Rough-Shifts*, but low overall scores 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, the pair with one high and one low coherence score, the experts agreed with the high Centering coherence measure, but they debated about the essay with the low Centering coherence measure. For that essay, one expert noted that "coherence comes and goes" while another found coherence in the essay's "chronological organization of examples" (a notion beyond the domain of Centering Theory). In the second case, the pair with low coherence scores but differing overall scores, our experts' judgments confirmed the *Rough-Shift* coherence measure. In the third case, the pair with one essay containing one fully developed point and another essay with several undeveloped points, our experts specifically identified both the coherence and the development aspects as determinants of the essays' scores. In general, our experts felt that the use of an automated coherence measure would be a valuable instructional aid.

To understand how an automated coherence measure could be applied to writing instruction, consider that effective writing teachers must spend much time providing students with diagnostic feedback about the specific strengths and weaknesses in their essays (Kukich 2000). A student who receives a low score wants to understand precisely where specific problems occurred in the essay. The *Rough-Shift* measure could be used to direct students' attention to specific sentences in need of improvement. It could pinpoint locations within an essay where topic discontinuities, i.e., rough shifts, appear to occur. For example, an interactive tutorial system could highlight segments containing *Rough-Shift* transitions. It could illuminate 'choppy' topic and focus chains within the text of an essay by using different colors for noun phrases playing the roles of Cb's and Cp's. Supplementary instructional comments could guide the student into revising the relevant section paying attention to topic discontinuities.

9 Remaining issues and future work

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. Further research is required to evaluate what the effect that the use of such expressions has on textual coherence, compared with simpler entities such as *John* or *the newspaper.* [15] 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 (25) and (26) below the coherence link is established by the semantics of the verb 'changes' and the noun 'change'.

(25) Many software companies changed their policy.

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

Within the Centering framework it is possible to treat these cases as cases where an utterance has no Cb. As correctly pointed out by a reviewer, it is in fact possible that an utterance has no Cb.[16] This is always the case, for example, when the utterance is discourse initial, but utterances with no Cb may also be found segment medially (Poesio, Cheng, Henschel, Hitzeman, Kibble and Stevenson 2000). In these cases, the Centering literature is unclear as to what the effect of 'No Cb' is on the computation of transitions. Elsewhere, as well as in this study, we consider a discourse medial utterance with 'No Cb' equivalent to an utterance whose Cb is different from the previous utterance and the Cb of the current utterance is different

[15] 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.

[16] It is also possible that an utterance has no Cp. For example, it is hard to identify the Cp in the following example taken from one of the students' essays:

(1) Does "not possible" mean "not likely"?

from the Cp of the current utterance. This means that a discourse medial utterance with no Cb yields a *Rough-Shift* transition as there is no link to establish coherence between two consecutive utterances, either by continuing on the previous center or promoting a new center.

On the other hand, discourse deixis and nominalizations are qualitatively different and we would like to distinguish them from cases with no Cb. Unlike cases with no Cb to establish a link between two utterances, discourse deictic expressions and nominalizations do establish a coherence link between the current utterance and the previous discourse. 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. In the original formulation of the Centering Model, discourse centers are defined as discourse constructs that establish a link between the current, previous and subsequent discourse. Discourse centers are semantic objects, not 'words, phrases, or syntactic forms' (Grosz, Joshi and Weinstein 1995). It was later shown that in most cases discourse centers can conveniently be mapped to syntactic forms ((Brennan, Friedman and Pollard 1987; Kameyama 1985), inter alia), but as we see in the case of nominalizations, for example, this mapping is not always trivial. To return to our system, even if we forced it 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 synonymy or more complex inferencing.

An alternative approach to this problem might be to undertake a study using the theme-rheme constructs (e.g., Daneŝ (1974).[17] It is possible that the coherence link to be identified in this case is the progression from the rhematic part in (25), i.e., *changed their policy*, to the thematic part of (26), i.e., *this change.*[18] Both approaches to this problem require a major research effort. Since this issue currently 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.

Our study prescribes a route for several future research projects. Some, such as the need to improve on fully automated techniques for noun phrase/discourse entity identification and coreference resolution, are essential for converting this measure of local coherence to a fully automated procedure. Others, not explicitly discussed here, such as the status of discourse deictic expressions, nominalization resolution, and global coherence studies are fair game for basic, theoretical research.

[17] Strube and Hahn (1996; 1999) recast Centering notions in terms of Daneŝ's (1974) trichotomy between given information, theme, and new information. The Cb(U_i), the most highly ranked element of Cf(U_{i-1}) realized in U_i, corresponds to the element which represents given information. The Cp(U_i) corresponds to the theme of U_i. The rhematic elements of U_i are the ones not contained in U_{i-1}. Note, however, that Strube and Hahn's implementation of this approach is restricted to nominal phrases in the exclusion of verb phrases or larger clausal parts which could instantiate the theme or the rheme of a sentence. Also, their proposal for ranking the elements of the Cf list according to their information status is motivated by German and it is not clear if it applies to English.

[18] We would like to thank an anonymous reviewer for pointing us to this direction.

Acknowledgements

We would like to thank Jill Burstein who provided us with the essay set and human and *e-rater* scores used in this study; Mary Fowles, Peter Cooper, and Seth Weiner who provided us with the valuable insights of their writing assessment expertise; Henry Brown who kindly discussed some statistical issues with us; Ramin Hemat who provided perl code for automatically computing Centering transitions and the *Rough-Shift* measure for each essay. We are grateful to Ellen Prince, Aravind Joshi and Alistair Knott for useful discussions.

References

Richard Becker and John Chambers. 1984. *An Interactive Environment for Data Analysis and Graphics*. Wadsworth, Inc., Belmont CA.

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

Jill Burstein, Karen Kukich, Suzanne Wolff, Martin Chodorow, Lisa Braden-Harder, Mary Dee Harris, and Chi Lu. 1998. Automated essay scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada*, August.

Guy Carden. 1982. Backwards anaphora in discourse context. *Journal of Linguistics*, 18:361–387.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psycological Measurement*, 20:37–46.

Frantisek Daneŝ. 1974. Functional sentence perspective and the organization of the text. In F. Daneŝ, editor, *Papers on Functional Sentence Perspective*, pages 106–128. Prague: Academia.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Barbara Di Eugenio. 1998. Centering in Italian. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 115–137. Clarendon Press: Oxford.

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

Barbara Grosz and Candace Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, MIT Press, Cambridge*, pages 44–50.

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

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

Irene Guggenmoos-Holzmann. 1993. How reliable are chance-corrected measures of agreement? *Statistics in Medicine*.

Marti Hearst. 1994. Multiparagraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Accosiation for Computational Linguistics*.

Caroline Heycock and Anthony Krock. 1997. Inversion and equation in copular sentences. In *Workshop on (Pseudo)Clefts, ZAS, Berlin*.

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

Felicia Hurewitz. 1998. A quantitative look at discourse coherence. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, chapter 14, pages 273–291. Clarendon Press: Oxford.

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

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

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

Megumi 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.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (Student Session)*, pages 286–288.

Helena Chmura Kraemer. 1982. Kappa coefficient. In *Encyclopedia of Statistical Sciences.* New York: John Wiley & Sons.

Karen Kukich. 2000. Beyond automated essay scoring. In *IEEE Intelligent Systems*, pages 22–27. September/October 2000.

Susumu Kuno. 1972. Functional sentence perspective: A case study from Japanese and English. *Linguistic Inquiry*, 3:269–320.

Thomas Landauer. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

M. Maclure and W.C. Willett. 1988. Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126(2):161–169.

Eleni Miltsakaki. 1999. Locating topics in text processing. In *Computational Linguistics in the Netherlands: Selected Papers from the Tenth CLIN Meeting, (CLIN '99), Utrecht*, pages 127–138.

Eleni Miltsakaki. 2001a. Centering in Greek. In *Proceedings of the 15th International Symposium on Theoretical and Applied Linguistics, Thessaloniki.*

Eleni Miltsakaki. 2001b. On the interpretation of weak and strong pronominals in Greek. In *Proceedings of the 5th International Conference in Greek Linguistics, Sorbonne.*

Eleni Miltsakaki. 2002a. Effects of subordination on referential form and interpretation. In *Proceedings of the 26th Penn Linguistics Colloquium.* Penn Working Papers in Linguistics, Philadelphia.

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

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

Frederick Mosteller and John Tukey. 1977. *Data Analysis and Regression, a Second Course in Statistics.* Addison-Wesley, Reading MA.

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

Ellis Page. 1966. The imminence of grading essays by computer. *Phi Delta Kappan*, 48:238–243.

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

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

Rebecca 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.

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

Massimo Poesio, Hua Cheng, Renate Henschel, Janet Hitzeman, Rodger Kibble, and Rosemary Stevenson. 2000. Specifying the parameters of centering theory: a corpus-based evaluation using text from application-oriented domains. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong-Kong.*

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1972. *A Grammar of Contemporary English*. London:Longman.

Tanya Reinhart. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27(1):53–93.

Jeff Reynar. 1994. An automatic method of finding topic boundaries. In *Proceedings of the 32nd Annual Meeting of the Accosiation for Computational Linguistics (Student Session)*, pages 331–333.

M.E. Schreiner, Bob Rehder, Thomas Landauer, and Darrell 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.

Candace Sidner. 1979. Towards a computational theory of definite anaphora comprehension in English discourse. Technical Report No. AI-TR-537, Artificial Intelligence Laboratory, Cambridge:Massachusetts Institute of Technology.

Michael Strube and Udo Hahn. 1996. Functional Centering. In *Proceedings of the 34th Annual Conference of the Association for Computational Linguistics (ACL '96), Santa Cruz*, pages 270–277.

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

Linda Suri, Kathleen McCoy, and Jon DeCristofaro. 1999. A methodology for extending focusing frameworks. *Computational Linguistics*, 25(2):173–194.

Izumi Tanaka. 2000. Cataphoric personal pronouns in English news reportage. In *Proceedings of the 3rd International Conference of Discourse Anaphora and Anaphor Resolution Conference, DAARC 2000, Lancaster*, volume 12, pages 108–117.

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

Karen van Hoek. 1997. *Anaphora and Conceptual Structure.* Chicago: University of Chicago Press.

Marilyn Walker and Ellen Prince. 1996. A bilateral approach to givenness: A hearer-status algorithm and a Centering algorithm. In T. Fretheim and J. Gundel, editors, *Reference and Referent Accessibility*, pages 291–306. Amsterdam: John Benjamins.

Marilyn Walker, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of Centering. *Computational Linguistics*, 20(2):193–233.

Marilyn Walker, Aravind Joshi, and Ellen Prince. (eds). 1998. *Centering Theory in Discourse.* Clarendon Press: Oxford.

Marilyn 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.

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

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

Appendix

Score	File	Continue	Retain	Smooth-Shift	Rough-Shift	% of Rough-Shifts
6	1	5	4	1	3	23
6	2	6	4	-	1	9
6	3	5	1	2	1	10
6	4	5	2	3	-	0
6	5	7	-	1	2	20
6	6	3	2	1	4	40
6	7	3	-	-	-	0
6	8	7	4	1	3	20
						$< 25\%$
5	9	11	-	-	-	0
5	10	9	-	2	2	15
5	11	6	2	3	3	21
5	12	4	1	3	1	11
5	13	7	6	-	1	7
5	14	2	4	1	2	22
5	15	7	3	2	4	25
5	16	5	2	-	2	22
						$< 25\%$
4	17	1	2	3	4	40
4	18	3	-	2	4	44
4	19	1	-	1	2	50
4	20	1	-	-	5	83
4	21	1	1	1	2	40
4	22	-	-	1	2	66
4	23	4	4	4	1	7
4	24	9	-	2	-	0
						$> 40\%$
3	25	2	1	1	3	50
3	26	-	-	-	-	-
3	27	3	-	2	4	44
3	28	1	1	1	3	50
3	29	-	3	-	2	40
3	30	3	1	-	-	0
3	31	-	-	-	-	-
3	32	2	2	1	3	37
						$> 40\%$

Table 5. *Table with Centering transitions of 32 GMAT essays*

HUM	E-R	ROUGH	E(PRED)	E+R(PRED)
6	5	15	5.05	5.26
6	6	22	5.9921	5.9928
6	6	15	5.99	6.09
6	6	22	5.9921	5.9928
6	6	24	5.99	5.96
6	4	22	4.13	4.35
6	4	13	4.13	4.46
6	6	28	5.99	5.90
6	5	30	5.0577	5.0594
6	4	30	4.13	4.24
6	4	0	4.13	4.62
6	5	20	5.05	5.19
6	6	21	5.99	6.00
6	6	50	5.99	5.58
6	6	25	5.99	5.94
6	5	21	5.05	5.18
6	6	6	5.99	6.22
6	5	35	5.05	4.98
6	5	25	5.05	5.12
6	5	30	5.057	5.059
5	4	15	4.14	4.46
5	5	7	5.07	5.40
5	4	5	4.14	4.60
5	5	38	5.07	4.96
5	4	40	4.14	4.12
5	5	45	5.07	4.86
5	6	27	6.02	5.95
5	4	30	4.28	4.14
5	5	21	5.07	5.20
5	5	16	5.07	5.27
5	5	20	5.07	5.22
5	6	32	6.02	5.88
5	4	40	4.143	4.148
5	4	10	4.14	4.53
5	4	23	4.14	4.35
5	5	20	5.07	5.22
5	6	25	6.02	5.98
5	4	25	4.14	4.33
5	5	50	5.07	4.79
5	6	10	6.02	6.20
4	3	11	3.22	3.71
4	5	45	5.09	4.88
4	4	46	4.15	4.04
4	3	50	3.22	3.17
4	3	36	3.22	3.37
4	3	33	3.22	3.41
4	5	42	5.09	4.92
4	3	50	3.22	3.17
4	4	36	4.15	4.18
4	4	40	4.15	4.13

Table 6. Table with the human scores (HUM), the *e-rater* scores (E-R), the *Rough-Shift* measure (ROUGH), the (jackknifed) predicted values using *e-rater* as the only variable, E(PRED), and the (jackknifed) predicted values using the *e-rater* and the added variable ROUGH, E+R(PRED). 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.

HUM	E-R	ROUGH	E(PRED)	E+R(PRED)
4	3	11	3.22	3.71
4	3	75	3.22	2.79
4	4	38	4.15	4.16
4	3	62	3.22	3.00
4	4	12	4.15	4.53
4	4	40	4.15	4.13
4	5	48	5.09	4.84
4	3	9	3.22	3.74
4	3	81	3.22	2.69
4	3	100	3.22	2.34
3	3	55	3.24	3.11
3	4	30	4.16	4.28
3	4	81	4.16	3.59
3	4	42	4.16	4.11
3	3	50	3.24	3.18
3	3	66	3.24	2.96
3	3	42	3.24	3.30
3	2	40	2.30	2.50
3	3	75	3.24	2.83
3	3	40	3.24	3.33
3	3	78	3.24	2.78
3	3	62	3.24	3.02
3	2	55	2.30	2.29
3	2	30	2.30	2.64
3	3	?	3.29	?
3	5	45	5.11	4.91
3	3	80	3.24	2.75
3	2	37	2.30	2.54
3	3	75	3.24	2.83
3	2	50	2.30	2.36
2	2	67	2.32	2.14
2	2	67	2.32	2.14
2	4	78	4.17	3.68
2	3	67	3.25	2.97
2	3	41	3.25	3.33
2	2	?	2.32	?
2	1	67	1.37	1.30
2	2	20	2.32	2.84
2	2	42	2.32	2.50
2	2	50	2.32	2.39
1	2	50	2.35	2.41
1	2	0	2.35	3.29
1	1	67	1.42	1.35
1	3	71	3.26	2.95
1	3	57	3.26	3.12
1	0	100	0.44	-0.03
1	1	85	1.42	1.09
1	1	67	1.42	1.35
1	2	57	2.35	2.31
1	1	0	1.42	2.48

Table 7. (Continued from Table 4)Table with the human scores (HUM), the *e-rater* scores (E-R), the *Rough-Shift* measure (ROUGH), the (jackknifed) predicted values using *e-rater* as the only variable, E(PRED), and the (jackknifed) predicted values using the *e-rater* and the added variable ROUGH, E+R(PRED). 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.

Score	File	Continue	Retain	Smooth-Shift	Rough-Shift
6	1	10	3	4	3
6	2	4	1	3	5
6	3	6	4	4	4
6	4	9	2	5	3
6	5	8	3	1	3
6	6	8	1	5	4
6	7	13	1	3	5
6	8	5	2	2	4
6	9	10	3	6	8
6	10	15	2	5	7
6	11	23	1	3	2
6	12	10	0	3	7
6	13	4	6	2	0
6	14	6	3	2	3
6	15	11	2	3	4
6	16	4	0	9	13
6	17	5	1	0	2
6	18	10	3	5	4
6	19	7	0	5	4
6	20	14	0	2	7
5	21	12	3	0	1
5	22	7	5	7	2
5	23	3	3	2	5
5	24	5	1	4	6
5	25	6	1	5	10
5	26	4	2	2	3
5	27	6	0	3	4
5	28	10	1	0	2
5	29	10	1	2	3
5	30	5	1	5	3
5	31	7	3	2	3
5	32	13	1	3	8
5	33	9	0	0	1
5	34	4	3	3	3
5	35	5	0	3	2
5	36	8	6	5	5
5	37	3	0	0	1
5	38	3	1	4	3
5	39	12	11	4	3
5	40	7	0	1	6

Table 8. *Table with Centering transitions for essay scores 5 and 6.*

Score	File	Continue	Retain	Smooth-Shift	Rough- Shift
4	41	2	3	2	2
4	42	3	0	0	7
4	43	7	0	1	5
4	44	3	0	3	6
4	45	4	1	2	4
4	46	2	2	0	2
4	47	1	1	2	3
4	48	1	2	3	6
4	49	7	1	3	7
4	50	8	1	3	7
4	51	5	3	0	1
4	52	0	0	1	3
4	53	5	0	3	5
4	54	0	1	2	5
4	55	6	0	1	1
4	56	4	0	1	2
4	57	6	4	3	12
4	58	4	4	2	1
4	59	0	2	0	9
4	60	0	0	0	5
3	61	2	2	0	5
3	62	2	4	1	3
3	63	2	0	0	9
3	64	6	0	2	6
3	65	1	0	3	4
3	66	0	1	1	4
3	67	4	0	0	3
3	68	4	1	0	3
3	69	0	0	1	3
3	70	1	1	1	2
3	71	2	0	1	11
3	72	0	1	3	5
3	73	2	1	1	5
3	74	6	1	0	7
3	75	7	1	1	4
3	76	0	0	0	0
3	77	4	5	3	10
3	78	0	0	1	4
3	79	3	0	2	3
3	80	0	1	3	12

Table 9. *Table with Centering transitions for essay scores 3 and 4.*

Score	File	Continue	Retain	Smooth-Shift	Rough-Shift
2	81	0	0	1	2
2	82	0	1	0	2
2	83	0	0	2	7
2	84	2	2	0	8
2	85	5	0	2	5
2	86	0	0	0	0
2	87	1	0	1	4
2	88	0	2	2	1
2	89	2	1	1	4
2	90	1	1	1	3
1	91	1	0	1	2
1	92	2	2	1	0
1	93	0	0	1	2
1	94	1	1	0	5
1	95	2	1	0	4
1	96	0	0	0	2
1	97	0	0	1	6
1	98	1	0	0	2
1	99	3	0	3	8
1	100	1	0	0	0

Table 10. *Table with Centering transitions for essay scores 1 and 2. Note that the counts in Tables (8), (9), and (10) are based on the earlier Cf ranking rule proposed in Brennan, Friedman and Pollard 1987.*