The Role of Centering Theory's Rough-Shift in the Teaching and Evaluation of Writing Skills

Eleni Miltsakaki
University of Pennsylvania
Philadelphia, PA 19104 USA
elenimi@unagi.cis.upenn.edu

Karen Kukich
Educatinal Testing Service
Princeton, NJ 08541 USA
kkukich@ets.org

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 ETS's *e-rater* essay scoring system to examine whether local discourse coherence, as defined by a measure of Rough-Shift transitions, might be a significant contributor to the evaluation of essays. Our positive results indicate that Rough-Shifts do indeed capture a source of incoherence, one that has not been closely examined in the Centering literature. 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 in pronominal-free text.

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 scores as frequently as two human scores match each other (Burstein et al., 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 et al., 1983, inter alia). 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 and suggests a route for exploring Centering Theory's practical applicability to writing evaluation and instruction.

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

2 The *e-rater* essay scoring system

One goal of automatic essay scoring systems such as *e-rater* is to represent the criteria that human experts use to evaluate essays. The writing features that *e-rater* evaluates were specifically chosen to reflect scoring criteria for the essay portion of the Graduate Management Admissions Test (GMAT). These criteria are articulated in GMAT test preparation materials at http://www.gmat.org. In *e-rater*, 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 local 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 of features.

3 Overview of Centering

A synthesis of two different lines of work (Joshi and Kuhn, 1979; Joshi and Weinstein, 1981) and (Sidner, 1979; Grosz, 1977; Grosz and Sidner, 1986) yielded the formulation of Centering Theory as a model for monitoring local focus in discourse. The Centering model was designed to account for those aspects of processing that are responsible for the difference in the perceived coherence of discourses such as those demonstrated in (1) and (2) below (examples from Hudson-D'Zmura (1988)).

(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 common assumptions in Centering as discussed in the literature (e.g., Walker et al. (1998)). We present the assumptions and modifications we made for this study in Section 6.1.

4.1 Discourse segments and entities

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. It is precisely this distinction between looking back in the discourse with the Cb and projecting preferences for interpretations in the subsequent discourse with the Cp that provides the key element in computing local coherence in discourse.

4.2 Centering transitions

Four types of transitions, reflecting four degrees of coherence, are defined in Centering. They are shown in transition ordering rule (1). The rules for computing the transitions are shown in Table 1.

(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.

	Cb(Ui)=Cb(Ui-1)	Cb(Ui)≠Cb(Ui-1)
Cb(Ui)=Cp	Continue	Smooth-Shift
Cb(Ui)≠Cp	Retain	Rough-Shift

Table 1: Table of transitions

4.3 Utterance

In early formulations of Centering Theory, 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. Based on crosslinguistic studies, Miltsakaki (1999) defined the utterance as the traditional 'sentence', i.e., the main clause and its accompanying subordinate and adjunct clauses constitute a single utterance.

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 crosslinguistically. Kameyama (1985) and Brennan et al. (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 arbitrary plural pronominals.

(3)Revised rule for the ranking of forward-looking centers: SUBJ>IND. OBJ>OBJ>OTHERS>QIS, PRO-ARB.

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.[3]

5 The role of Rough-Shift transitions

As mentioned briefly earlier, the Centering model includes one more rule, the Pronoun Rule given in (4).

(4) Pronoun Rule: If some element of Cf(Ui-1) is realized as a pronoun in Ui, then so is the Cb(Ui).

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-Shifts 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, inter alia). 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. Our data consist of student essays whose degree of coherence is under evaluation and therefore cannot be assumed. Using students' paragraph marking as segment boundaries, we 'centered' 100 GMAT essays. The average length of these essays was about

[3] But see also Di Eugenio (1998) for the treatment of complex NPs in Italian.

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

Table 2: Distribution of forms over Rough-Shifts

250 words. In the next section we show that Rough-Shift transitions provide a reliable measure of *incoherence*, correlating well with scores provided by writing experts.

One of the crucial insights was that, in our data, the incoherence detected by the Rough-Shift measure is not due to violations of the Pronominal Rule or infelicitous use of pronominal forms in general. In Table 2, we report the results of the distribution of forms over Rough-Shift transitions. Out of the 211 Rough-Shift transitions, found in the set of 100 essays, in 195 occasions the Cp was a nominal phrase, either definite or indefinite. Pronominals occurred in only 16 cases of which 6 cases instantiated the pronominals 'we' or 'you' in their generic sense. Table 2 strongly indicates that student essays were not incoherent in terms of the processing load imposed on the reader to resolve anaphoric references. Instead, the incoherence in the essays was due to discontinuities in students' essays caused by their introducing too many undeveloped topics within what should be a conceptually uniform segment, i.e. their paragraphs. This is, in fact, what Rough-Shift picked up.

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 resoluion 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 succession of multiple entities. That is, Rough-Shifts are the result of absent or 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 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 human raters with respect to the corresponding requirements described in the instructions for human raters. We observed that essays with higher scores 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 appendix). 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.

6.1 Centering assumptions and modifications

Utterance. Following Miltsakaki (1999), we assume that the each utterance consists of one main clause and all its subordinate and adjunct clauses.

Cf ranking. We assumed the Cf ranking given in (3).

A modification we made involved the status of the pronominal *I*. [4]We observed that in low-scored essays the first person pronominal *I* was used extensively, normally presenting personal narratives. However, personal narratives were unsuited to this essay writing task and were assigned lower scores by expert readers. 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.

Lack of Fit Source	DF	Sum of Squares	Mean Square	F-Ratio
Lack of Fit	71	53.55	0.75	1.30
Pure Error	24	13.83	0.57	Prob>F
Total Error	95	67.38	0.23	
Max RSq 0.94				
Parameter Estimates Term	**Estimate**	**Std Error**	**t-Ratio**	**Prob>$\|t\|$**
Intercept	1.46	0.37	3.92	0.0002
E-RATER	0.80	0.06	11.91	<.0001
ROUGH	-0.013	0.0041	-3.32	0.0013
Effect Test Source Nparm	**DF**	**Sum of Squares**	**F-Ratio**	**Prob>F**
E-RATER 1	1	100.56	141.77	<.0001
ROUGH 1	1	7.81	11.01	0.0013

Table 3: Regression

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

6.2 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 Cb and all 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. As output, the 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.

7 Study results

In the appendix, 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 HUMAN (HUM) column contains the essay scores given by human raters and the EARTER (E-R) column contains the corresponding score assigned by the *e-rater*. Comparing HUMAN 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 that 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. The 'Estimate' cell contains the coefficients assigned for each variable. The coefficient for ROUGH is negative, thus penalizing occurrences of Rough-Shifts in the essays. The t-test ('t-ratio' in Table 3) for ROUGH has a highly significant p-value ($p<0.0013$) for these 100 essays suggesting that the added variable ROUGH can contribute to the accuracy of the model. The magnitude of the contribution indicated by this regression is approximately 0.5 point, a reasonalby sizable effect given the scoring scale (1-6). Additional work is needed to precisely quantify the contribution of ROUGH. That would involve incorporating the ROUGH variable into the building of a new *e-rater* model and comapring the results of the new model to the original *e-rater* model.

As a preliminary test of the predictability of the model, we jacknifed the data. 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. We repeated the procedure using both variables. The predicted values for ERATER alone and ERATER+ROUGH are shown in columns PrH/E and PrH/E+R respectively in Table 4. In comparing the predictions, we observe that, indeed, 57 % of the predicted values shown in the PrH/E+R column are better approximations of the HUMAN scores, especially in the cases where the ERATER's score is discrepant by 2 points from the HUMAN score.

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 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, 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.

The advantage of the Rough-Shift metric over other quantified components of the *e-*

rater is that it can be appropriately translated into instructive feedback for the student. In an interactive tutorial system, segments containing Rough-Shift transitions can be highlighted and supplementary instructional comments will guide the student into revising the relevant section paying attention to topic discontinuities.

9 Future work

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.

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 Aravind Joshi and Alistair Knott for useful discussions.

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.

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

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.

S. Hudson-D'Zmura. 1988. *The Structure of Discourse and Anaphor Resolution: The Discourse Center and the Roles of Nouns and Pronouns*. Ph.D. thesis, University of Rochester.

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.

E. Miltsakaki. 1999. Locating topics in text processing. In *Proceedings of Computational Linguistics in the Netherlands (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.

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.

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 and E. Prince. 1995. A bilateral approach to givenness: A hearer-status algorithm and a Centering algorithm. In T. Fretheim and J. Gundel, editors, *Reference and Referent Accessibility*. Amsterdam: John Benjamins.

M. Walker, A. Joshi, and E. Prince (eds). 1998. *Centering Theory in Discourse*. Clarendon Press: Oxford.

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.

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	5.05	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

HUM	E-R	ROUGH	PrH/E	PrH/E+R
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 4: Table with the human scores (HUM), the *e-rater* scores (E-R), the Rough-Shift measure (ROUGH), the (jacknifed) predicted values using *e-rater* as the only variable (PrH/E) and the (jacknifed) 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.