Computational Linguistics and Chinese Language Processing
Vol. 14, No. 2, June 2009, pp. 181-204

An Exploratory Application of Rhetorical Structure Theory to Detect Coherence Errors in L2 English Writing: Possible Implications for Automated Writing Evaluation Software

Sophia Skoufaki*

Abstract

This paper presents an initial attempt to examine whether Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) can be fruitfully applied to the detection of the coherence errors made by Taiwanese low-intermediate learners of English. This investigation is considered warranted for three reasons. First, other methods for bottom-up coherence analysis have proved ineffective (e.g., Watson Todd *et al.*, 2007). Second, this research provides a preliminary categorization of the coherence errors made by first language (L1) Chinese learners of English. Third, second language discourse errors in general have received little attention in applied linguistic research. The data are 45 written samples from the LTTC English Learner Corpus, a Taiwanese learner corpus of English currently under construction. The rationale of this study is that diagrams which violate some of the rules of RST diagram formation will point to coherence errors. No reliability test has been conducted since this work is at an initial stage. Therefore, this study is exploratory and results are preliminary. Results are discussed in terms of the practicality of using this method to detect coherence errors, their possible consequences about claims for a typical inductive content order in the writing of L1 Chinese learners of English, and their potential implications for Automated Writing Evaluation (AWE) software, since discourse organization is one of the essay characteristics assessed by this software. In particular, the extent to which the kinds of errors detected through the RST analysis match those located by *Criterion* (Burstein, Chodorow, & Leachock, 2004), a well-known AWE software by Educational Testing Service (ETS), is discussed.

* Graduate Institute of Linguistics, National Taiwan University
E-mail: sophiaskoufaki@ntu.edu.tw

[Received August 10, 2009; Revised December 6, 2009; Accepted December 22, 2009]

Keywords: Automated Writing Evaluation, Discourse Organization, Coherence Errors, Rhetorical Structure Theory.

1. Introduction

Research findings indicate that English language learners produce various kinds of discourse errors in their writing, such as inductive patterns (e.g., Kaplan, 1966) and inappropriate coordination (e.g., Soter, 1988). However, the discourse errors of second language (L2) learners of English have not been examined in detail partly because at least some of them are more difficult to detect than other kinds of errors (e.g., syntactic, spelling). This paper describes an initial attempt to examine whether Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) can be fruitfully applied to the detection of the coherence errors made by Taiwanese low-intermediate learners of English. In particular, this paper reports on a pilot study where 45 written samples from the LTTC English Learner Corpus, a Taiwanese learner corpus of English currently under construction, were analysed according to RST. It is hoped that this pilot study will provide some preliminary indication of the viability of this approach to coherence error detection.

The results of this analysis will also serve as a preliminary list of coherence errors which may prove typical or not in further large-scale studies of this kind. A categorization of second language (L2) English coherence errors in general and of the coherence errors of particular learner populations has not been provided yet by applied linguists. Therefore, this pilot study is warranted because of its possible utility for research on English L2 discourse and the instruction of writing in English as an L2.

Another aim of this study is to examine whether the most frequent of the errors detected through the RST analysis can be located by *Criterion*, a well-known AWE software by the Educational Testing Service (ETS). Automated Writing Evaluation (AWE) software such as *Criterion* (e.g., Burstein, Chodorow, & Leachock, 2004) and *My Access!* (e.g., Vantage Learning, 2007) locate and give diagnostic feedback only for a limited number of discourse errors. This issue has been pointed out by the computational linguists involved in the creation of AWE software (e.g., Higgins, Burstein, Marcu, & Gentile, 2004), but no study has been conducted with specific English learner populations to examine what discourse errors should be added to the inventory of discourse errors currently located via AWE software. Being a pilot study, the study reported here does not purport to fill this research gap but only to provide an initial step towards this goal.

In the following two sections, this paper will offer further information on the motivation of this study. Then, it will offer some background information on RST. Third, it will provide an overview of the LTTC English Learner Corpus and will describe the data and method of the study. Fourth, it will describe findings from a qualitative and quantitative perspective. Fifth,

these results will be discussed in relation to a) whether RST analysis seems a viable method for coherence error detection, b) which factors seem to affect the coherence errors located in the data, c) whether results indicate inductive order patterns and d) how much they overlap with the coherence errors that can be located via *Criterion*. The paper will end with a summary of conclusions and directions for future research.

2. RST and Discourse Coherence Error Detection

It is difficult to reliably identify coherence errors because readers of the same text may form different interpretations of the coherence relations among elements of the text (Mann & Thompson, 1988). Therefore, a bottom-up method of coherence error detection should be used so that coherence errors will be identified as reliably and objectively as possible.

RST was chosen first because the output of other methods of locating coherence breaks, such as topical structure analysis and genre analysis in Watson Todd *et al.* (2007), has been shown to have little relationship with English teachers' judgments. Second, strong correlations have been found between RST analyses which show that a text is coherent and subjective judgments that a text is coherent (Taboada & Mann, 2006a). Finally, RST has not been applied to the location of coherence errors (Higgins, Burstein, Marcu, & Gentile, 2004: 185), so an evaluation of its application for this purpose is interesting from a methodological perspective.

3. Discourse Coherence and L1 Chinese Learners of English

Given the paucity of discourse error tagging in learner corpora (Díaz-Negrillo & Fernández-Domínguez, 2006) and the sparse research on discourse errors by learners of English, this pilot study aims to provide a preliminary categorization of discourse errors in the writing of low-intermediate Taiwanese learners of English. This list of errors will be supplemented and refined through further research.

L1 Chinese learners of English make similar discourse errors to learners with other native tongues, but there have also been claims for typical L1 Chinese errors. However, these claims have not been examined sufficiently through quantitative methods. Therefore, the pilot study reported in this paper also partly functions as a preliminary quantitative test for one of these claims. This claim is that the paragraphs and essays of L1 Chinese learners of L2 English have an inductive rather than deductive order. It has been claimed that these learners present the main point of their writing only at the end of a paragraph or essay, whereas in L1 English writing the main point is presented first (e.g., Kaplan, 1966; Matalene, 1985).

The claim for the use of an inductive order only by L1 Chinese learners of English (and not by native speakers of English) has been challenged. For example, Scollon and Scollon (1995) used ethnomethodology to show that inductive and deductive patterns both exist in the speech of

both native speakers of English and native speakers of Chinese. The only difference between the two languages is that these patterns are used for different pragmatic purposes. However, their analysis relates only to spoken discourse, so one cannot draw any conclusions about the existence of inductive patterns in written native English. This research gap is filled by Chen (2008). In a quantitative study, he found, among other things, that the minority of the native speakers of English preferred essays written with an inductive rather than deductive pattern and nearly half of them preferred paragraphs written in an inductive rather than a deductive order. This finding indicates that inductive patters can be used in written English but they are more acceptable in paragraphs rather than in essays. Finally Mohan and Lo (1985) review Chinese writing textbooks and analyse Classical Chinese texts to show that the deductive pattern is the most usual and prescribed essay writing pattern in Chinese[1].

From a theoretical perspective, if the RST analysis of the texts in the pilot study can point to instances of inductive order, the controversial issue of whether the English discourse of L1 Chinese learners is characterized by inductive order will be able to be examined in more detail in later research. Moreover, if the present study indicates that inductive-order errors occur frequently in the data, this may be seen as a preliminary indication that AWE software should try to detect and categorize as errors cases of inductive content order.

4. Discourse Errors and *Criterion*

The pilot study reported here is also motivated by one of the criticisms made about AWE software, that is, that the effectiveness of AWE software should not be tested only through "*a posteriori* statistical validation" but also through an "*a priori* investigation of what *should* be elicited by the test before its actual administration" (Weir, 2005: 17). In other words, high levels of agreement in the grades assigned to essays between human judges and software should not be the only criterion for software evaluation; the kinds of errors which are located by software should also match those located by human judges. Such concerns are warranted for practical reasons as well, since it has been shown that learners can fool AWE software, that is, they can get high scores although the content of their essays is inadequate (Herrington & Moran, 2001; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Ware, 2005). Therefore, if AWE software is designed so as to locate the errors that a human judge would locate, wrong essay

[1] Controversy also exists over the cause of inductive patterns whenever they are found in the writing of L1 Chinese learners of English. For example, one possible reason is the influence from L1 rhetorical structure, as contrastive rhetoric theorists claim (eg., Chen, 2001; Kaplan, 1966; Matalene, 1985). Another is the lack of relevant or useful feedback and instruction from teachers (e.g., Gonzales, Chen, & Sanchez, 2001; Mohan & Lo, 1985). Yet another possible reason is the inability to properly structure an essay not only in the L2 but also in the L1 because one has not reached the right developmental stage in his/her writing ability (e.g., Mohan & Lo, 1985).

evaluations will be prevented.

To examine this issue, this section of the paper will summarize the kinds of discourse errors located by *Criterion*[2]. Then, section 9.3 will compare them with the discourse errors located through the RST analysis in the present study. The rationale is that any discrepancies between the two lists of errors should warrant large-scale empirical work testing whether these discrepancies really exist.

The main discourse errors which are located by *Criterion* are those of absence or insufficient number of discourse structures considered necessary in expository and argumentative essays, which are the input of this software. This is a valuable feature because no other AWE software has it (Burstein, 2009: 15). These structures are introductory information which forms the background for the rest of the essay ('Introductory Material'), the statement which expresses the opinion of the writer ('Thesis Statement'), the main point(s) made by the writer ('Main Point'), the statement(s) which support(s) each main point ('Support'), and the conclusion ('Conclusion'). For example, if a learner has not included a thesis statement in his or her essay, the software is likely to locate this error and inform the learner about it.

Apart from the aforementioned discourse-structure tags, the creators of *Criterion* had initially used separate tags for cases where learners had written a title for their essay, for opening and closing salutations in essays in letter format, and for content which could not be tagged with any of the other tags. These tags occurred infrequently, so such cases were lumped under the tag 'Other' (Burstein, 2009: 15; Bustein, Marcu, & Knight, 2003: 33). However, this practice obscures the number of times when the software could not categorize structures through any of the existing labels. This problem could be important because perhaps structures could not be labeled by the software because they violated the usual order of discourse structures (that is, Introductory Material, Thesis, Main Points, Conclusion), an error which should occur whenever information is ordered unusually in an essay. This possibility is likely because in *Criterion* one of the modules used to identify the discourse structures in essays is the 'global language model'. It predicts the sequence of discourse elements in an essay by seeing how well the predictions which stem from a 'local language model' - which predicts which discourse structure is likely to appear after two sections which have already been tagged as specific kinds of discourse structures - fit a final-state grammar manually created by the software creators (Burstein, Marcu, & Knight, 2003: 36).

As we have seen in section 3 of this paper, inductive, rather than deductive, content order has been claimed to characterize the writing of Chinese L1 learners of English; therefore, the

[2] *Criterion*, rather than *My Access!*, was chosen because the research reports on the latter do not give enough information about its workings for its discourse organization evaluation function to be assessable.

software's inability to locate such errors could lead to its low efficacy whenever L1 Chinese learners order their essay content inductively.

A related problem is that because most parts of the software leading to discourse evaluation in *Criterion* are probabilistic, they rely on the most frequent patterns found in essays which were commented on and graded by human graders. This means that errors which did not occur often could not be identified by the software. For example, thesis statements which are scattered in the essay instead of being expressed through one or more adjacent clauses cannot be identified. In an evaluation of the latest version of *Criterion*, the discourse structures that could not be categorized in the training data were 13% of this data (Burstein, Marcu, & Knight, 2003: 36).

One apparent problem with this software is that it presupposes that the essays written will have one paragraph as an introduction, one as a conclusion, and three paragraphs in between, each expressing and supporting one main point. This is an expected structure for a short essay, but this assumption in the software also means that it cannot locate these discourse elements in essays which have fewer or more paragraphs. For example, the writing samples from the lower-intermediate GEPT examination, which form the data for the current study, would not be able to be evaluated by *Criterion* since most of them are one-paragraph long.

Criterion also assesses how balanced the development of an essay is by calculating the proportion of the words in each discourse structure as compared to the total number of words in an essay. This measure seems useful, but it is a crude way of examining degree of development because the length of a structure in terms of its constituent words does not necessarily correlate with how rich it is in content. For example, some learners could repeat the same point in order to meet the required word limit.

Criterion can also decide whether an essay is off-topic or not and also whether content in one or more of the main-point paragraphs is off-topic.

This overview of the discourse errors which *Criterion* can identify shows that it can detect important errors of discourse organization (namely, whether the usual discourse elements occur) and content (namely, whether all or part of an essay is off-topic). This overview has also shown that this software cannot assess essays for discourse coherence, since it cannot identify cases of unusual ordering of content or content which is irrelevant to a specific segment of a text rather than to the essay topic. Recently, research has been conducted with the aim to improve *Criterion* so that it can produce more fine-grained feedback about discourse organization (Higgins, Burstein, Marcu, & Gentile, 2004), but it is still in a preliminary stage. Findings regarding the assessment of coherence inside discourse segments were not encouraging because the criterion used – whether a sentence was related to at least one other sentence in the same discourse segment – was met in the vast majority (92.81%) of sentences (Higgins, Burstein, Marcu, &

Gentile, 2004: 190).

5. A brief Introduction to RST

In their review of theoretical work on RST, Taboada and Mann (2006a: 425) give a simple definition of RST: "RST addresses text organization by means of relations that hold between parts of a text. It explains coherence by postulating a hierarchical, connected structure of texts, in which every part of a text has a role, a function to play, with respect to other parts in the text." The connections which are posited between parts of a text and which show the function of each 'part of text' in the text are called 'coherence relations'. Coherence relations show the function that the analyst thinks that the writer intended each 'part of text' to have in relation to other parts of text.

Some units are called 'nuclei' and others 'satellites'. In RST jargon, nuclei are units of analysis which are necessary parts of a text and satellites are units of analysis which modify the meaning of the nuclei. The main idea of a text needs the nuclei to be put across but if the satellites were deleted, the same main idea, more or less, would be expressed.

For example, the analyst will say that there is an elaboration coherence relation between two units of analysis, if (s)he thinks that the author wishes that the reader recognize the satellite as providing additional information for the nucleus. Figure 1 shows an extract from a paragraph from the LTTC English Learner Corpus. The second and third clauses are linked through the relationship of 'joint' because one is added to the other and jointly modify the first sentence by elaborating its meaning ('elaboration').

[Your teacher may tell you lots of ways to keep your eyes from nearsightedness.] [Such as keep thirty centimeters from your eyes to the table,] [and not to read books when it's dark.]

Figure 1. Extract from a sample paragraph from the LTTC English Learner Corpus illustrating the 'elaboration' coherence relation; each unit of analysis appears within square brackets.

As mentioned above, the coherence relations in a text are usually presented in a hierarchical structure. Figure 2 shows the structure of the extract in Figure 1. The software used to produce it is the RST Annotation Tool by Daniel Marcu[3], which is an improvement on Marc O'Donnell's RSTTool[4]. In RST diagrams, coherence relations are indicated by arrows. An arrow starts from a satellite and points to a nucleus. However, there are also some coherence relations which link units of the same kind. The relation 'joint' is such a 'multinuclear' relation.

[3] This software was downloaded from http://www.isi.edu/licensed-sw/RSTTool/index.html.

[4] This software can be downloaded from http://www.wagsoft.com/RSTTool/section2.html.

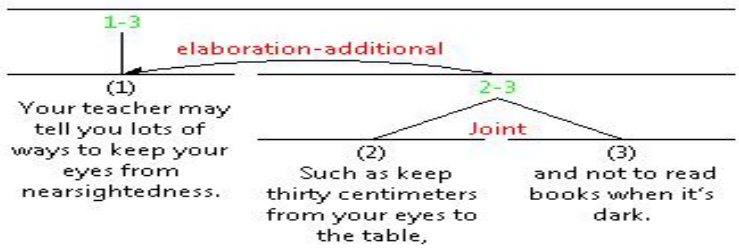

Figure 2. RST diagram indicating the coherence relations in the extract presented in Figure 1.

As mentioned earlier, the analyst chooses the coherence relation which seems to have the function that the writer intended each 'part of text' to have in relation to other parts of text. There are certain constraints on the analyst's choice of a coherence relation, but this paper will describe only one of them because, although they guided RST data analysis, the rest are not directly related to the method of this study. The constraint which helped to form the method of this project is that each text should have the structure of a coherence-relation schema. Such a schema is an abstract representation of coherence relation diagrams. The analyst tries to fit a whole text into one schema and to fit sub-schemas under this schema. Figure 3 shows the schemas which have been posited by Mann and Thompson (1987, 1988).

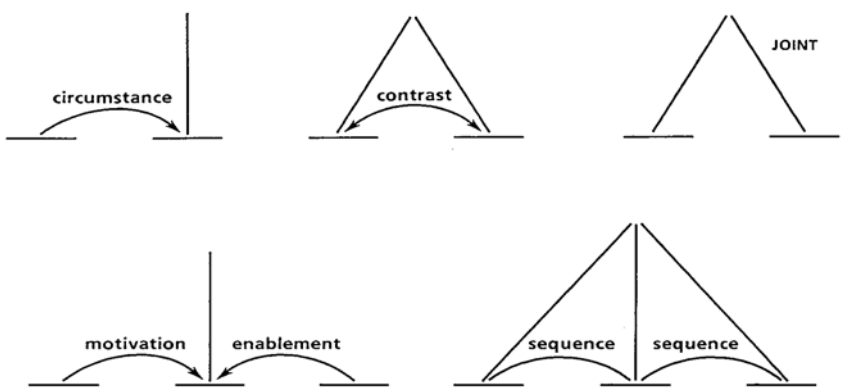

Figure 3. Schemas posited by Mann and Thompson (1987, 1988); figure taken from Mann and Thompson (1987:7).

The aforementioned schema application constraints have some consequences for the location of coherence errors. Since all these requirements must be met for a text to be considered coherent in RST, their violations indicate coherence errors. Therefore, coherence errors are expected to be indicated by diagrams which

a) do not comply with the structure of any schema,
b) include sub-diagrams which do not comply with the structure of any schema, or
c) include schemas which share units of analysis ('crossed dependencies').

This conclusion leads us to the rationale of this study: each kind of coherence error will be indicated by one of these abnormalities in the diagram. By listing the abnormalities which characterize each kind of coherence error, texts can later be tagged for coherence errors in a principled way.

6. Data

The data are 45 paragraphs written by Taiwanese lower-intermediate learners of English in Writing Task 2 of the Intermediate General English Proficiency Test (GEPT) examination, a language proficiency examination administered by the LTTC, a language testing company in Taiwan. In this task, test-takers are asked to write a 120-word paragraph. These files form part of the written section of the LTTC English Learner Corpus, which is currently under construction[5]. The corpus will consist of language samples by Taiwanese learners of English who have sat the GEPT. In the current, first phase of corpus construction, 2,000 written-production and 400 oral-production samples from the Intermediate GEPT examination have been processed.

In order to examine coherence errors in paragraphs written on more than one topic, the 45 paragraphs were equally distributed across topics. Topics were presented to test-takers in Chinese. Two of these topics are questions about personal preferences (favorite food and idol, respectively) and the third asked test-takers to explain why many elementary-school children in Taiwan are nearsighted and to propose effective ways of preventing nearsightedness.

To ensure that the data that would be analysed would vary in terms of coherence error types, samples were equally distributed across score bands in each topic. In other words, in each topic five files had low scores (ranging from 1 to 2), five files had medium scores (ranging from 2.5 to 3.5) and five had high scores (ranging from 4 to 5).

7. Method

The method involved the analysis of the aforementioned paragraphs by the author using the RST Annotation Tool software.

The units of analysis were defined in the same way as in the tagging of 385 documents of

[5] This corpus is compiled under the supervision of Professor Hintat Cheung, the director of the Graduate Institute of Linguistics at NTU. The co-directors are Professor Zhao-Ming Gao, from the Department of Foreign Languages and Literatures at NTU, and Professor Siaw-Fong Chung, from the Department of English at National Chengchi University. I am the postdoctoral research associate working on the project. The other project members are two PhD students, Ms Sally Chen and Ms Chi-Yi Wu, and the research assistant and administrator, Ms Su-Mei Chen. In the academic year 2008-9, the research assistant and administrator was Ms San-Ju Lin.

American English selected from the Penn Treebank (Carlson & Marcu, 2001). Broadly speaking, clauses were the units of analysis, except when they were complements of prepositions and verb objects. However, because the tagset that Carlson, Marcu and their collaborators used was specific to the nature of the texts which they analysed (that is, Wall Street Journal articles), I preferred to use the more neutral coherence relation categories by Bill Mann[6]. Since I combined the units of analysis from the Penn Treebank corpus and Bill Mann's categories, I had to compromise the unit-of-analysis segmentation when the units of analysis warranted a coherence relation which was not among those in Bill Mann's list. This happened when the coherence relation of 'attribution' was posited by Carlson and Marcu to link speech and thought verbs with their complements. In these cases, I considered the verb and its complement clause as one unit of analysis.

As the analysis of the texts was progressing, it became obvious that Bill Mann's list of relations could not cover all the coherence relations in the text, so they were supplemented with eight relations from the tagset by Carlson, Marcu and their collaborators (Carlson & Marcu, 2001). These additional coherence relations were: 'same-unit', 'comment', 'conclusion', 'topic-shift', 'manner', 'explanation-argumentative'.

8. Results

8.1 Qualitative Results

Table 1 summarizes the coherence breaks indicated by the main abnormalities found in the RST diagrams.

Table 1. Abnormalities found in the RST diagrams of the 45 data paragraphs and the coherence breaks indicated by them.

Diagram abnormalities	*Coherence breaks indicated by diagram abnormalities*
Dangling units of analysis	Irrelevant content Incomprehensible content 'Self-sufficiency'
Crossed dependencies	Although a sub-diagram has already been formed for one part of the text, a coherence relation arises between another text part and a unit which is a member of the first sub-diagram
Unexpected relation	Motivation
Relations occurring in unexpected parts of a diagram	Inductive content order

[6] These are the original categories posited by Mann and Thompson (1987, 1988) with some additions and can be found at this website: http://www.wagsoft.com/RSTTool/RSTDefs.htm.

Dangling units of analysis constitute the first kind of RST diagram abnormality. Clauses or larger elements which seem unrelated to the content of the rest of the text are unexpectedly linked to it through a coherence relation. Such dangling units indicated irrelevant content most of the time. There was one case where I left a unit dangling because it was impossible to understand it. Finally, there was one instance of a self-sufficient clause, which explained why the writer liked a specific foreign food in a postscriptum. This error can be categorized as one where the learner was unclear about the layout which (s)he was expected to use.

Figure 4 gives an example of a dangling unit with irrelevant content.

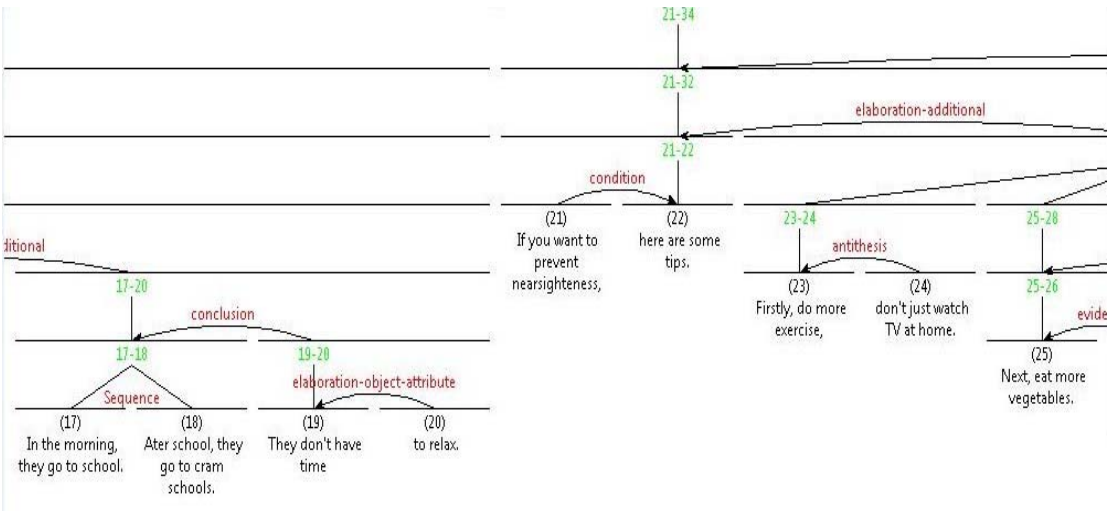

Figure 4. Extract from a diagram where the structure consisting from units 21-34 is dangling because it is irrelevant to preceding text.

This extract comes from a paragraph written on the topic about the nearsightedness of elementary school children in Taiwan. Since the topic asked test-takers to propose effective methods of preventing nearsightedness in general, the advice which the writer gives to the reader in the sub-diagram consisting of units 21 to 34 is irrelevant.

An example of crossed dependencies cannot be illustrated diagrammatically because the RST Annotation Tool automatically corrects such abnormalities in a diagram. However, one can consider the coherence relations among the units in the extract in Figure 5. This figure shows the first lines written in a paragraph on the favorite exotic food topic. In this figure, the units are numbered for ease of reference to them in the discussion that follows.

1. [Taiwan is a special country.] 2. [We can eat a lot of foods from other countries.] 3. [They are gathered in this small island.] 4. [Like Japan, America, Tailand and more.]

Figure 5. Extract from a paragraph on the favorite exotic food topic; each unit of analysis appears within square brackets and the number of each unit of analysis precedes it.

Unit 3 restates information given in unit 2, so 3 is the satellite and 2 the nucleus of a 'restatement' coherence relation. Together, they express a result which stems from the fact that Taiwan is a special country, expressed in unit 1. Therefore, units 2 and 3 together form the satellite of a 'result' coherence relation, where 1 is the nucleus. Unit 4 exemplifies the countries whose food the Taiwanese can eat in Taiwan, so it is the satellite of an 'elaboration' coherence relation and 2 is the nucleus. This coherence relation is problematic because unit 4 intrudes in the sub-diagram which has already been formed by units 2 and 3.

Unwarranted coherence relations constitute the next RST diagram abnormality. The only such coherence relation which was found in the pilot was 'motivation'. It is a coherence relation between a nucleus and a satellite, the latter of which offers a reason why the reader should do something which is expressed in the former. This relation is found in argumentative discourse (Azar, 1999) and not in expository and narrative discourse, which the GEPT test-takers were expected to produce. Figure 6 gives two examples of this error in an extract from a paragraph on the nearsightedness of elementary students in Taiwan.

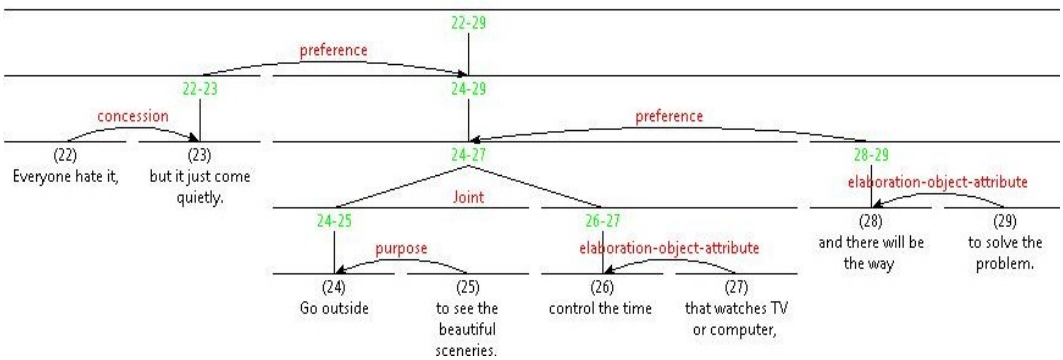

Figure 6. Extract from a diagram where a) the structure consisting of units 22-23 is the satellite in a 'motivation' coherence relation and the structure consisting of units 24-29 is its nucleus and b) the structure consisting of units 28-29 is the satellite in a 'motivation' coherence relation and the structure consisting of units 24-27 is its nucleus.

In units 22 and 23, 'it' refers to nearsightedness. These units jointly form a sub-diagram which serves as the satellite in a 'motivation' relation because they give a reason why someone should do the actions described in the units 24-29.[7] Units 28 and 29 have the same function for units 24-27, so they are the satellite in a 'motivation' relation as well.

Finally, coherence relations in inappropriate parts of a text are the last RST diagram

[7] The relations connecting units 22-23 to units 24-29 and units 28-29 to 24-27 are called 'preference' in this diagram only because the relation 'motivation' is not in the list of coherence relations in the RST Annotation Tool. Throughout the RST analysis of the data, the tag 'preference' was too stand for 'motivation'.

abnormality. These coherence relations are acceptable if they occur in the right parts of a text but there were cases where their location was inappropriate and indicated inductive content order. The 'conclusion' coherence relation indicates a relation where the satellite is a reasoned judgment, inference, necessary consequence or final decision. For example, a student explained why Taiwanese elementary pupils are nearsighted by giving the example of what happened to her younger brother and concluded that "playing video games and watching television too much may be closely related to the cause of elementary students' nearsightedness problem." The 'background' coherence relation usually appears in introductions or briefly in later parts of a text but when students use it extensively in the main body of a text, it may lead to inductive content.

It should be noted that although the RST analysis yielded a wealth of diagram problems which indicate coherence errors in the data, some coherence errors did not show up as problems in the RST diagrams. In other words, the aforementioned diagram abnormalities are not enough to pinpoint all the coherence errors in the data. There are cases where the writer inappropriately addresses the reader but this does not lead to a structural error in the diagram and cases where a topic sentence is missing or scattered in different parts of the text without affecting the diagram. Therefore, the intuition of the error tagger is always necessary for the location of coherence errors.

8.2 Quantitative Results

The variety of coherence errors which were indicated in the preceding qualitative analysis would not be meaningful in this study if it were not supplemented with an analysis aiming to see which errors are the most frequent. The rationale is that those errors which seem to occur often in the data may warrant further investigation in later, large-scale studies. However, these results should be interpreted with caution because they are based on an RST analysis which has been conducted by only one person and only once. In other words, they are based on data which have not been checked for this validity and reliability. Moreover, the number of writing samples analysed is small, so the descriptive statistics which will be presented here are far from statistically reliable. For this reason, inferential statistics have not been conducted on the data.

As it has been mentioned in the overview of the qualitative results, dangling structures usually indicated irrelevant content. However, there was also one case where I could not link a structure to a preceding sub-diagram because this structure was incomprehensible and another because it appeared in a post-scriptum. Because these two errors occurred only once each, I have excluded them from the calculations which resulted in the figures in Table 2. In this table, because written samples varied in terms of their length, the number of occurrences of dangling structures was divided by the total number of units of RST analysis in each sample. Thus, the frequency of this diagram abnormality was normalized in a way appropriate to the way texts

were analysed. The last column is a coarser estimation of the frequency of dangling structures per topic because it is the count of the texts which included at least one dangling structure.

Table 2. Irrelevant content instances across topics according to the RST analysis of 45 paragraphs.

Topic	Cumulative 'dangling' structures normalized per RST units of analysis	Mean 'dangling structures' normalized per RST units of analysis	Writing samples with at least one dangling structure; percentage of texts per topic is given within parentheses
Nearsightedness	0.377	0.021	5 (33.33%)
Idol	0.059	0.004	1 (6.66%)
Exotic food	0.111	0.012	2 (13.33%)

All three measures of frequency agree with each other in that in the 'nearsightedness' topic there are more dangling structures than in the other topics and that the 'exotic food' topic contains more dangling structures than the 'idol' topic. This finding can be seen as indicating that topic affects the occurrence of irrelevant content. Especially in terms of the last frequency measure, it is impressive that in one topic one third of the samples contained irrelevant content. All the frequencies are small, but it should be kept in mind that the maximum number of words in this task was only 120 words. In other words, the short word length created few 'opportunities' for irrelevant content to occur.

It was interesting to examine whether these differences in the frequency of dangling structures also seem related to the score band (low, medium, or high) under which the samples fall. In Table 3 below, the cumulative percentages of the dangling structures are presented in terms of essay topic and score band.

Table 3. Cumulative percentage of 'dangling' structures per topic and score band.

Score band	Essay topic		
	Nearsightedness	Idol	Exotic food
Low score	37.92%	0%	50%
Mid score	38.65%	100%	0%
High score	23.42%	0%	50%
Total percentage	100%*	100%	100%

The total number from the percentages in this column is 99.99% because these numbers are rounded. The exact total number is 100%.

The breakdown of samples which contain dangling structures in the nearsightedness topics is as expected, since one would expect that learners with low and mid scores would be more likely to include irrelevant content in their writing than the high-performing learners. The data from the other two topics is more complicated, since all cases of irrelevant content in the idol topic occurred in the middle-score paragraphs and half of them in the low- and the other half in the high-score paragraphs in the food topic. However, this finding can be easily explained by the very few occurrences of dangling structures for the idol and food topics. There was only one occurrence of a dangling structure in the idol topic and it was in a middle-score paragraph and there were only two occurrences in the food topic, one in a low- and the other in a middle-score paragraph.

In sum, results on dangling structures show that this type of RST diagram abnormality indicates irrelevant content and that the frequency of such errors depends on the essay topic, at least in the writing of these low-intermediate Taiwanese learners of English.

Coherence errors stemming from crossed dependencies are likely to be very rare since this data contains only one such error.

As explained in the previous section, the coherence relation of 'motivation' was unexpected because it normally occurs in argumentative text types whereas the essay topics were expository. This coherence relation occurred only in the paragraphs written on the 'nearsightedness' topic. This finding is congruent with the previous finding that irrelevant content made manifest by dangling structures was much more frequent in the nearsightedness than in the other texts. Indeed, it seems that there is some interrelation between dangling structures and the existence of a 'motivation' relation in nearsightedness texts, as shown in Table 4.

Table 4. Total and mean number of Motivation relation instances in the paragraphs written on the Nearsightedness topic according to the RST analysis and percentage of paragraphs which included both at least one 'motivation' relation and at least one 'dangling' structure.

Cumulative instances of 'motivation' coherence relation normalised per RST unit of analysis	Mean instances of 'motivation' relation normalized per RST units of analysis	Percentage of paragraphs with 'motivation' relation which also have 'dangling' structures
0.112	0.007	66.67%

In terms of to the coherence errors due to the occurrence of a coherence relation in an inappropriate part of a paragraph, Table 5 presents the same kinds of normalized data as Table 2 but for the inappropriate occurrences of the 'background' coherence relation.

Table 5. Inappropriate uses of the 'background' coherence relation across topics according to the RST analysis of 45 paragraphs.

Topic	Cumulative inappropriate uses of the 'background' coherence relation normalized per RST units of analysis	Mean inappropriate uses of the 'background' coherence relation normalized per RST units of analysis	Number of writing samples with at least one instance of an inappropriate use of the 'background' coherence relation; percentage of texts per topic is given within parentheses
Nearsightedness	0.059	0.004	1 (6.67%)
Idol	0.184	0.012	3 (20%)
Exotic food	0	0	0 (0%)

As it can be seen, the majority of cases occur in the idol topic, so it seems that the occurrence of such errors also depends on topic. To see whether there was a score-band effect as well, in Table 6 below, the cumulative percentages of the dangling structures are presented in terms of essay topic and score band.

Table 6. Cumulative percentage of cases of 'background' coherence relation per topic and score band (there were no cases in the 'exotic food' topic).

Score band	Essay topic	
	Nearsightedness	Idol
Low score	0%	0%
Mid score	0%	42.83%
High score	100%	57.17%
Total percentage	100%	100%

This table indicates that the 'background' coherence relation occurred in the wrong part of the text for paragraphs which achieved medium and high scores. This finding may not be significant in the 'nearsightedness' topic since this error was only found in one paragraph, but it seems to be more important in the 'idol' topic since this error occurred in one fifth of these paragraphs.

The last coherence error indicated by the RST analysis is the use of the 'conclusion' coherence relation in an inappropriate part of the text. This error occurred only twice and only in two middle-score paragraphs, so it seems that this error occurs rarely. Moreover, it occurred only in the 'nearsightedness' topic, so this error is also possibly due to a topic effect.

9. Discussion

9.1 RST Analysis as a Means to Coherence Error Detection

The results of this pilot study indicate that different kinds of abnormalities in RST diagrams built on writing samples of low-intermediate GEPT test-takers indicate various coherence errors. In particular, dangling units and unexpected coherence relations in the diagrams are indications of irrelevant content. Coherence relations in inappropriate parts of the text indicate inductive content order. Finally, the crossed dependencies indicate local coherence errors because they apply to coherence relations within rather than across sub-diagrams. Consequently, this method of textual analysis seems promising. However, as it has been mentioned in section 8.2 above, this method cannot detect all coherence errors that a human analyst can. Moreover, it is labor-intensive, so it may be impractical to use. Therefore, if this method proves effective – that is, if in a large-scale study inter- and intra-judge reliability are high and the agreement between the coherence errors located by the RST analysis and the judgments of language teachers and/or native speakers is high – it should be used by skilled analysts and only when fine-grained analyses of coherence errors are desirable.

9.2 Coherence Errors by Low-intermediate Taiwanese Learners of English

As mentioned in section 3, this pilot study also aimed to examine which coherence errors are made by a specific population of English language learners, namely, low-intermediate Taiwanese learners of English. As mentioned above, errors of irrelevant content, inductive content order, local coherence errors due to crossed dependencies, use of an inappropriate coherence relation (i.e., 'motivation') and the occurrence of coherence relations 'background' and 'conclusion' in inappropriate parts of a text have been detected via the RST analysis. However, as mentioned in section 3, inappropriate addresses to the reader were not detected through the RST analysis. These addresses are inappropriate because the topics were expository and addresses to the reader are common in argumentative writing. Cases where a topic sentence is missing or scattered in different parts of the text could also not be detected through the RST analysis.

The frequency of the coherence error types which could be detected through RST in the data should be considered with caution given the small number of paragraphs, their short word length, and the fact that the analysis was not checked for inter- and intra-judge reliability. Keeping this caveat in mind, one can note that the 'dangling structure' RST diagram abnormality is the most frequent one. The second most frequent RST abnormality was the inappropriate use of the 'background' coherence relation. The third most frequent RST

abnormality was the unwarranted occurrence of the 'motivation' coherence relation.[8] All the other RST diagram abnormalities occurred so infrequently that it seems that they are unlikely to occur frequently in a large-scale study. There was only one case of crossed dependencies and only two cases of inappropriate use of the 'conclusion' coherence relation.

As indicated in the quantitative analysis of the data in section 8.2, all coherence errors located in the data seem to vary depending on topic. Most dangling structures occur in paragraphs on the 'nearsightedness' topic and most cases of inappropriate use of the 'background' coherence relation occur in paragraphs on the 'idol' topic; the 'motivation' coherence relation and the inappropriate use of the 'conclusion' coherence relation occur only in paragraphs on the nearsightedness topic. These indications of topic effects – which could be due to topic content, phrasing or other topic characteristics – point to the need to investigate the occurrence of coherence errors for this population further. Analysing larger numbers of data and writing samples from a larger variety of topics will be able to indicate which coherence errors are frequent irrespective of writing topic and, therefore, warrant more attention from English language teachers and AWE software.

The quantitative analysis by both score band and topic for the two most frequent errors, namely, 'dangling structures' and the inappropriate use of the 'background' coherence relation, showed a different trend for each of these errors. For the 'dangling structure' error, this analysis was done only for the answers to the 'nearsightedness' topic because, contrary to the other two topics, it received a number of such errors big enough for this data analysis to be meaningful. The breakdown of error numbers according to score bands was as expected since most 'dangling structures' occurred in the low- and mid-score paragraphs. In terms of the inappropriate use of the 'background' coherence relation, only the answers to the 'idol' topic received enough answers for the analysis per score band to be meaningful. Here, the results were different from those for the 'dangling stucture' errors because most such errors occurred in the paragraphs which had received high scores. Moreover, the second most error-populated score band was the mid one and no such error occurred in the low-score band. As mentioned in section 8.1, 'dangling structure' errors indicate irrelevant content and the inappropriate use of the 'background' coherence relation indicates inductive content order. The contrasting aforementioned results between the two error types can be explained through a consideration of the literature on the criteria used in essay marking. Research indicates that in L1 essay grading, the focus is on discourse organization whereas in L2 essay grading the focus is on syntactic and lexical errors (e.g., Breland & Jones, 1982; Gonzáles, Chen, & Sanchez, 2001). This fact may

[8] These frequency comparisons were made according to all measures used in this study (that is, cumulative occurrences normalized by RST units of analysis, mean occurrences normalized per RST units of analysis, and number of writing samples with at least one occurrence).

explain why some students received mid- and high-grades although their paragraphs were partly organized inductively. Another possible explanation is that these paragraphs formed part of students' answers to the level of the GEPT examination aimed to low-intermediate learners of English; markers considered that more local errors should weigh more in the marking than discourse errors. Both these possible explanations indicate the need for an examination of errors other than discourse errors in the data, so that the gravity of these errors in each score-band can be compared against that of the discourse errors. Therefore, to reach a conclusion about whether the discourse errors located often in this pilot study occur frequently in the writing of low-intermediate Taiwanese learners of English in general, we do not only need to repeat this analysis with more writing samples and two analysts, but also the writing samples should be tagged for other errors as well. The development of an error tagging system for the LTTC English Learner Corpus is currently under way.

Another central aim of the study was to examine whether inductive order errors are frequent in the writing of low-intermediate Taiwanese learners of English. The relatively frequent occurrences of the 'background' coherence relation in inappropriate parts of a paragraph indicate that these learners make such errors. However, the concentration of such errors in the paragraphs written on the 'idol' topic indicates the possibility of topic effects. As mentioned above, a large-scale study with samples written on a larger variety of topics is necessary to measure the frequency of inductive order errors and their relation to topic effects.

9.3 Coherence Errors Detected via RST Analysis and *Criterion*

As mentioned in section 4, a secondary aim of this pilot study is to examine the extent to which the errors located through the RST analysis can also be located via *Criterion*. This section will examine this issue by considering each coherence error separately.

The most frequent kind of RST diagram abnormality in the data of the present study is 'dangling structures'. In all except two cases, dangling structures indicate irrelevant content, so if this finding proves valid through a large-scale study, AWE software should be able to locate irrelevant-content errors. It is unclear to me whether *Criterion* would be able to categorize such cases as off-topic. The first method used in *Criterion* to locate off-topic essays and segments is through comparisons of the vocabulary used in the essays used for training the software to score and give feedback on a specific topic. The second method is a comparison between the proportion of times in which a word is used in a variety of topics and that where a word is used in a specific topic. The third method does not require training data and relies on a comparison of the vocabulary in the essay prompt and in the essays (Burstein, 2009: 8-12). In segments like the dangling one in Figure 4, topic-related vocabulary is used, so such segments would probably not be categorized as irrelevant to the topic by any of these methods. However, one should keep in mind that the high concentration of irrelevant content errors in the paragraphs written on the

nearsightedness topic indicates the possibility of topic effects on the occurrence of this error. Large-scale studies are necessary to clarify whether this error occurs often in the writing of low-intermediate L1 Chinese learners of English irrespective of topic-related factors.

The inappropriate use of the 'background' coherence relation is the second most frequent coherence error located through the RST analysis of the data. As explained earlier, it indicates inductive content order. Inductive content order cannot be detected by *Criterion*. However, the high concentration of this error in paragraphs written on the 'idol' topic may mean that this finding is just due to a topic effect and may not occur across topics. If large-scale studies indicate this, *Criterion* and other AWE software would not need to detect this error. Moreover, Chen's (2008) finding that paragraphs with inductive content order were acceptable by half the English native speakers participants in his study may mean that inductive order content should not be considered an error and calls for further examination of what makes inductive content order more or less acceptable for a native English speaker.

The next most frequent coherence error detected through the RST analysis is the unwarranted occurrence of the 'motivation' coherence relation. It is unclear whether the methods which detect off-topic content would be able to locate unwarranted instances of the 'motivation' relation in an essay. However, in any case, the fact that this relation occurred only in paragraphs written on the 'nearsightedness' topic may indicate a strong topic effect and large-scale studies which manipulate topic characteristics should be conducted to examine whether such coherence errors occur often enough and across topics to warrant the creation of AWE software which can detect them.

The inappropriate use of the 'conclusion' coherence relation occurred only twice. Its very low frequency probably means that AWE software would not need to locate this coherence error. As for the inappropriate use of the 'background' coherence relation, it cannot be detected by *Criterion*. This error was the second most frequent one, but again, most of its instances occurred in paragraphs written on only one topic, the 'idol' one. This finding warrants large scale studies which will examine whether such errors occur frequently across topics for this learner population.

The local coherence error caused by crossed dependencies occurred only once in the data. There is some controversy over whether such diagrammatic structures should be considered erroneous, because it has been claimed that crossed dependencies occur in the productions of native speakers as well (Wolf & Gibson, 2004, 2005). Therefore, such errors probably do not warrant further investigation or location through AWE software.

10. Conclusion

The main finding of this study is that an RST analysis of short texts written by low-intermediate L1 Chinese learners of English can provide detailed information about coherence errors. Nevertheless, the study presented here is only a pilot, so it was conducted on a small number of texts and only by one researcher. Therefore, this paper mainly serves as an index of research questions that need to be addressed through further research. As mentioned above, the inter- and intra-judge reliability of the analysis remains to be tested and the frequency of each kind of coherence error located needs to be measured through the analysis of larger numbers of texts.

The results of this pilot study strongly indicate the possibility of topic effects on coherence error occurrence. Therefore, further examination is also necessary to examine whether the topic effects on coherence errors occur when more samples are analysed. Moreover, texts written on topics which vary in terms of various dimensions (e.g., text type associated with a topic, how clearly the topic question explains what the essay structure should be) should be examined to examine whether certain kinds of topics lead to certain kinds of coherence errors. If these topic effects are confirmed through further research, attempts should be made to explain them. This research would be beneficial for AWE design, since if the topic-related factors shown to influence these errors could be detected by AWE software, essay scoring and feedback would be refined.

References

Azar, M. (1999). Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory. *Argumentation*, 13(1), 97-114.

Burstein, J. (2009). Opportunities for Natural Language Processing Research in Education. In Gebulkh, A. (Ed.), *Springer lecture notes in computer science* (Vol. 5449, pp. 6-27). Springer: New York, NY.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online Writing Evaluation service. *AI Magazine*, 25(3), 27-36.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. In Harabagiu, S. & Ciravegna, F. (Eds.), *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 18(1), 32-39.

Carlson, L. & Marcu, D. (2001). *Discourse Tagging Manual.* ISI Tech Report ISI-TR-545.

Chen, J.P. (2001). *Markedness in intercultural discourse: a study of Chinese EFL students' discourse patterns.* PhD thesis, Guangdong University of Foreign Studies, in ERIC, RIE June 2001.

Chen, J.P. (2008). An investigation into the preference for discourse patterns in the Chinese EFL learning context. *International Journal of Applied Linguistics*, 18(2), 188-211.

Díaz-Negrillo, A. & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, 19, 83-102.

González, V., Chen, C-Y., & Sanchez, C. (2001). Cultural Thinking and Discourse Organizational Patterns Influencing Writing Skills in a Chinese English-as-a-Foreign-Language (EFL) Learner. *Bilingual Research Journal*, 25(4), 417-442.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, 63(4), 480-499.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In Dumais, S., Marcu, D. & Roukos, S. (Eds.), *Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL) 2004: Main Proceedings* (pp. 185-192). Boston, MA: Association for Computational Linguistics.

Kaplan, R. (1966). Cultural thought patterns in intercultural education. *Language Learning*, 16(1), 1-20.

Mann, W.C. & Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization* (No. ISI/RS-87-190). Marina del Rey, CA: Information Sciences Institute.

Matalene, C. (1985). Contrastive rhetoric: an American writing teacher in China. *College English*, 47(8), 789-808.

Mohan, B.A. & Lo, W.A.-Y. (1985). Academic writing and Chinese students: transfer and developmental factors. *TESOL Quarterly*, 19(3), 515-34.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stamping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2), 103-134.

Scollon, R. & Wong-Scollon, S. (1995). *Intercultural communication: A discourse approach.* Blackwell: Oxford, U.K. and Cambridge, MA, U.S.A.

Taboada, M. & Mann, W.C. (2006a). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8(3), 423-459.

Taboada, M. and Mann, W.C. (2006b). Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4), 567-588.

Vantage Learning. (2007). *MY Access! Efficacy Report***.** Newtown, PA: Vantage Learning. Retrieved on August 6 2009 from http://www.vantagelearning.com/docs/myaccess/myaccess.research.efficacy.report.200709.pdf.

Ware, P. (2005). Automated writing evaluation as a pedagogical tool for writing assessment. In A. Pandian, G. Chakravarthy, P. Kell, & S. Kaur (Eds.), *Strategies and practices for improving learning literacy* (pp. 174-184). Selangor, Malaysia: Universiti Putra Malaysia Press.

Watson Todd, R., Khongput, S., & Drasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing*, 12(1), 10-25.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York: Palgrave MacMillan.

Wolf, F. & Gibson, E. (2004). Representing Discourse Coherence: A Corpus-based Analysis. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (article no. 134), Geneva, Switzerland.

Wolf, F. & Gibson, E. (2005). Representing Discourse Coherence: A Corpus-based Analysis. *Computational Linguistics*, 31(2), 249-287.