# Disunity in Cohesion: How Purpose Affects Methods and Results When Analyzing Lexical Cohesion

**Stuart G. Towns** **Richard Watson Todd**
Department of Applied Linguistics
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
sgtowns@gmail.com irictodd@kmutt.ac.th

## Abstract

Lexical Cohesion is a commonly studied linguistic feature as it is easily identified from the surface of a text. However, the purposes for studying lexical cohesion are varied, and each purpose requires different methods. This study analyzes two short movie review texts for four different research purposes using lexical cohesion: text evaluation, text segmentation, text summarization, and text criticism. The analysis shows that these four different purposes produce very different results concerning the lexical cohesion of the two texts, suggesting that the apparently straightforward construct of lexical cohesion is actually complex.

## 1 Introduction

The purposes of text analysis research can be divided into two main categories: applications and descriptions. The difference between these two areas is that applications produce results that are useful to end users who are outside of the field of linguistics, while descriptions of language are used internally by the linguistic community (Sinclair, 2004a). Many text analysis applications created for those outside of linguistics use automated tools, and therefore they focus on features that can be identified and analyzed with computers. One linguistic feature that can be analyzed to varying degrees of success using computers is lexical cohesion, since lexical cohesion can be found in the surface features of text. The analysis of lexical cohesion has been used in many text analysis applications, such as discourse analysis (Morris & Hirst, 1991), automatic text summarization (Barzilay & Elhadad, 1999), text segmentation (Stokes, Carthy, & Smeaton, 2004), word sense disambiguation (Okumura & Honda, 1994), and evaluation of machine translations (Wong & Kit, 2012).

Lexical cohesion was defined by Halliday and Hasan (1976, p. 274) as "the cohesive effect achieved by the selection of vocabulary" and is one of five types of cohesion (the other four being reference, substitution, ellipsis, and conjunction). A cohesive text is held together by explicit relationships found in the lexis and grammar of the text. These lexico-grammatical relationships are called cohesive ties as they connect one sentence to another. Multiple ties can, in turn, be combined into longer lexical chains which can span large portions of the text.

Current technology can identify lexical cohesion ties and lexical chains of ties with varying degrees of accuracy. Some cohesive ties are very easy to identify, such as the exact repetition of a lexical unit in an adjacent sentence, while others can be more difficult to correctly identify, such as the relationship of a pronoun to a noun in a previous sentence. Hoey (1991) outlined six types of lexical cohesion which are ordered by ease of identification from easiest to most difficult, along with some examples in Table 1.

As stated earlier, lexical cohesion has been used in text analysis research for many different


28th Pacific Asia Conference on Language, Information and Computation pages 513–521

| Lexical Cohesion Type | Definition | Examples (Hoey, 1991) |
|---|---|---|
| Simple repetition | Repetition of a word (singular or plural) | bear/bear, bear/bears |
| Complex Repetition | Repetition of two lexical items with a common stem, but different parts of speech | historical/history, quoted/quotation |
| Simple Paraphrase | Where a lexical item can replace another lexical item without a change in meaning | volume/book, writings/works |
| Complex Paraphrase | Antonymy, or the presence (or absence) of two links creating a third link | hot/cold, writer/writings/author, teacher/(teaching)/instruction |
| Semantic Association | Superordinate, Hyponymic, Co-reference | bears/animals, scientists/biologists |
| Non-lexical repetition | Personal and demonstrative pronouns | he, she, it, they, this, that, these, those |

Table 1: Hoey's (1991) six types of lexical cohesion

purposes. This paper will look at four main purposes: text evaluation, text segmentation, text summarization, and text criticism. The first three of these can be analyzed using automated computerized tools, while the fourth is a qualitative analysis that is beyond the capabilities of today's computers.

These four purposes can be described as follows. The first purpose, text evaluation, especially of student writing, has often focused on the lexical cohesion of the text as a marker of the quality of the text, with the assumption being that features such as referential cohesion correlate with human evaluations of high quality text (Weston, Crossley, & McNamara, 2010). The second purpose, text segmentation, finds breaks in the text where there are no lexical chains. The lack of lexical chains in a span of text shows that the topic might have changed (Şimon, Gravier, & Sébillot, 2013). The third purpose, text summarization, tries to identify the important topics in the text in order to create a summary of the text. Lexical cohesion aids this task by showing which topics are repeated throughout the text (Barzilay & Elhadad, 1999). The fourth purpose, text criticism, looks at the lexical cohesion in a text and attempts to understand the meaning behind the lexical choices, for example to find metaphors in political speeches that support the speaker's public image (Klebanov, Diermeier, & Beigman, 2008).

A key issue for lexical cohesion analysis is that the unit on which the analysis is conducted differs depending on the purpose of the research. Each of the four purposes discussed in this paper investigate a different unit. For the first purpose, a text evaluation is an evaluation of the cohesiveness of the text as a whole, and therefore should be based on the entire text. This can be done, for example, by computing the average cohesion between all of the sentences in the text. Text segmentation is an attempt to segment the whole text into smaller units and therefore the analysis must be based on units that are smaller than the whole text, such as measuring the lexical cohesion between individual adjacent pairs of sentences. Text summarization is focused on the lexical items of the text in order to find the important concepts, so the cohesive lexical items take priority over the whole text itself. Text criticism is not only looking at the lexical choices made by the writer or speaker but also at the potential meaning behind these choices. Therefore, the unit of investigation can vary in length as needed.

Even though all of these purposes are using lexical cohesion as the subject of research, the results of the research may be very different. The purpose of this paper, then, is to illustrate how different purposes require different methods, and how

these methods can lead to very different results depending on the operationalization of lexical cohesion, whether it is lexical cohesion of a text as a whole, lexical cohesion between adjacent sentences, lexical cohesion chains, or lexical cohesion created through nearby items in the same semantic sets.

## 2 Methodology

The texts to be used for the lexical cohesion analyses in this study will be movie reviews. Eight movie reviews of Wes Anderson's *Moonrise Kingdom* were downloaded from the Internet. Four of the reviews were written by Pulitzer Prize movie reviewers while four were written by amateur movie review bloggers. These eight movie reviews were analyzed using Coh-Metrix, an automated web-based tool which was originally created to automatically analyze text for cohesion and readability (Graesser, McNamara, Louwerse, & Cai, 2004). It was found that two of the reviews, one written by a Pulitzer Prize winner and one written by a blogger, showed very similar high scores relative to the other six reviews on the Coh-Metrix indices for lexical cohesion, or what Coh-Metrix calls referential cohesion.

This study will analyze these two texts for the four research purposes mentioned above: text evaluation, text segmentation, text summarization, and text criticism. The text written by the blogger will be labeled Text 1 and the text written by the Pulitzer Prize winner will be labeled Text 2. Text 1 has 324 words and 14 sentences while Text 2 has 758 words and 19 sentences (for the full texts, see Luke, 2012 for Text 1 and Hornaday, 2012 for Text 2.)

Since each of the four research purposes focuses on a different aspect of the text, each one has its own methodology. For text evaluation, which looks at the cohesion of the text as a whole as a measure of the quality of the text, the first analysis tool that will be used is Coh-Metrix. The eight lexical cohesion indices in Coh-Metrix represent averages across the text of the scores of the lexical cohesion between pairs of sentences. A binary score of either 1 for cohesion or 0 for no cohesion is found for every pair of adjacent sentences as well as for every sentence compared to every other sentence in the text, and is then averaged to give one number for each index.

Another way to analyze the cohesion of the text as a whole is to consider the lexical cohesion chains. Averages can be computed for the whole text for metrics such as the number of lexical chains, chain length, and chain density (defined here as the number of lexical items in the chain divided by the number of sentences in the chain).

For text segmentation, it is desirable for the results to show where the text has topic breaks. Therefore, a unit smaller than the entire text should be analyzed. As in the first analysis, lexical cohesion will be identified in pairs of adjacent sentences, but it will be done using a moving window approach (Stokes, Carthy, & Smeaton, 2004) where the individual scores for sentence-pair lexical cohesion are computed. A topic break occurs when a sentence pair does not have any shared lexical cohesive items. The text will then be divided into segments at these topic breaks. The length of the segments and the number of segments will be compared between the two texts to find out if there are any differences between the lexical cohesion in each.

For text summarization, the analysis is attempting to find the important topics in the text. These important topics will occur frequently in lexical cohesion chains running through the text. Therefore, the analysis will focus on the lexical items that can combine to form topics by looking at the number of lexical items inside each chain and the length of the chain. The chains will be mapped to show how much of the text they cover, as well as the location of the lexical items inside the chains. The patterns created by the lexical cohesion chains can then be compared between the two texts.

| Coh-Metrix Lexical Cohesion (Referential Cohesion) Indices | Text 1 | Text 2 | Avg of 6 |
|---|---|---|---|
| Noun Overlap, Adjacent Sentences, Binary, Mean | .385 | .500 | .215 |
| Noun Overlap, All Sentences, Binary, Mean | .294 | .370 | .194 |
| Stem Overlap, Adjacent Sentences, Binary, Mean | .615 | .611 | .275 |
| Stem Overlap, All Sentences, Binary, Mean | .435 | .437 | .241 |
| Argument Overlap, Adjacent Sentences, Binary, Mean | .846 | .667 | .452 |
| Argument Overlap, All Sentences, Binary, Mean | .529 | .548 | .383 |
| Content Word Overlap, Adjacent Sentences, Proportional, Mean | .067 | .047 | .066 |
| Content Word Overlap, All Sentences, Proportional, Mean | .046 | .039 | .048 |

Table 2: Referential cohesion results from Coh-Metrix for Text 1 and Text 2

For text criticism, the purpose is to understand the meaning behind the important words in the text. This requires a qualitative analysis of the lexical cohesion chains as well as the words that are collocated with these chains.

Throughout this paper so far, and in many other studies, there has been no distinction made between the terms "text" and "corpus". However, mentioning this potential distinction might be helpful to describe the difference between the first three methods (text evaluation, text segmentation, and text summarization), and the fourth (text criticism). Viewing data as a corpus (as was done for the first three methods) implies that automated tools will be used to observe the data. The researcher must choose the appropriate tool or must create their own tool depending on the type of information that is desired. Viewing the data as a text, on the other hand, means that the analysis will be done in a similar fashion to a human reading the text (Sinclair, 2004b). The first three analyses view the movie review data as a corpus, and have used automated tools to analyze the data. The fourth analysis will take a more human approach, viewing the data as a text to be read and understood. In this fourth analysis, the words themselves are not as important as the implied meaning behind the words in the mind of the reader.

## 3 Results

The first research purpose that will be considered is text evaluation. For this purpose, texts in their entirety are analyzed to find an overall lexical cohesion score. This analysis was done using Coh-Metrix on all eight original movie review texts. It was found that two of the texts, which are labeled in this study as Text 1 and Text 2, had similar, high cohesion scores for many of the Coh-Metrix indices compared to the other six texts. For example, for the Coh-Metrix index "Stem Overlap, all sentences, binary, mean", Text 1 scored .435 and Text 2 scored a very similar .437. The average of the other six texts was much lower at .241. The results from the Coh-Metrix analysis for Text 1, Text 2, and the average of the other six texts are found in Table 2.

Another way to measure the cohesion of the text as a whole is to investigate the lexical cohesion chains that are in the text. There are several metrics related to lexical chains that can be found, as seen in Table 3. These numbers, in contrast to the ones in Table 2, show some major differences between the two texts. Text 2 has 36% more sentences than Text 1, but three times more lexical chains. This means that, on average, there are more cohesive lexical items in each sentence in Text 2.

In addition, the lexical chains in Text 2 are on average longer and less dense than the ones in Text 1. This means that the cohesive ties are more likely to span longer distances in Text 2 than in Text 1.

The lexical chain patterns also show a lot of difference between the two texts. Half of the lexical chains in Text 1 are two-sentence chains with just one cohesive tie. Text 2 on the other hand, has several chains with one tie that span four sentences.

The second research purpose considered was text segmentation. To segment the text, lexical cohesion can be used to find topic breaks. Wherever there is no cohesion between adjacent sentences, it may be a signal that the topic of the text has changed. By analyzing the two texts using a two-sentence moving window, it can be seen that the two texts would be segmented very differently.

The segmentation of Text 1 is straightforward. It can be divided into three segments, as seen in Table 4. Segment 1 covers sentences 1-5, Segment 2 covers sentences 6-10, and Segment 3 covers the remaining sentences 11-14. The segmentation of Text 2 is more complicated. It can be divided into five segments. The first segment covers

| | **Text 1** | **Text 2** |
|---|---|---|
| Text length | 14 sentences | 19 sentences |
| # of Lexical Chains | 7 | 22 |
| Avg. Chain Length | 4.0 sentences | 6.0 sentences |
| Longest Chain | 13 sentences | 18 sentences |
| Avg. Chain Density | 81% | 44% |
| Most common pattern | 2-sentence chains with 1 tie (100% density) | 4-sentence chains with 1 tie (25% density) |

Table 3: Whole-text cohesion chain metrics

| Sentence | | | | Sentence | | | |
|---|---|---|---|---|---|---|---|
| 1 | film | words | Anderson | 1 | house | | |
| 2 | medium | words | he | 2 | house | created | Anderson |
| 3 | film | | he | 3 | artisanal | create | Anderson |
| 4 | M.K. | | | 4 | damp canvas | | |
| 5 | it | | | 5 | | | |
| 6 | start | story | | 6 | | house | |
| 7 | start | story | Anderson | 7 | Hayward | house | |
| 8 | M.K. | | Anderson | 8 | Hayward | | |
| 9 | M.K. | summer | Anderson | 9 | Sam,Suzy | M.K. | |
| 10 | | summer | | 10 | Sam,Suzy | M.K. | adults |
| 11 | film | | | 11 | Suzy | | grown-ups |
| 12 | it | | | 12 | players | | |
| 13 | film | | | 13 | plays | | |
| 14 | film | | | 14 | play | film | |
| | | | | 15 | solemnity | films | Anderson |
| | | | | 16 | solemnity | | Anderson |
| | | | | 17 | | | Anderson |
| | | | | 18 | | | Anderson |
| | | | | 19 | | | |

Table 4: Two-sentence moving window cohesion showing text segmentation

sentences 1-4. Then, sentence pairs 4-5, and 5-6 do not have any lexical cohesion, which means that there is a sentence-long break between the first two segments. The next three segments of sentences 6-11 and 12-18 are straightforward. The last sentence does not have any lexical cohesion with the sentence before it, so it is counted as the fifth segment.

The third research purpose was text summarization. To accomplish this, lexical cohesion chains can be analyzed to find the important topics in the text. The methodology here is different than what was done for text segmentation above in that the focus is on words rather than sentences. These lexical cohesion chains can span multiple sentences, and the lexical items do not necessarily have to be in adjacent sentences. Looking at the lexical cohesion chains that were analyzed for the first research purpose of text evaluation, the frequency of the lexical cohesive units within the chains can be seen in Tables 5 and 6. Text 1 has 7 lexical cohesion chains and Text 2 has 22 lexical cohesion chains.

The lexical chains that appear in a text can point to the important topics of the text. There are two ways that a summarization might be done. If the desired result is simply a noun phrase (i.e., a single short topic for the whole text), then the most frequent lexical items in the longest chains might form this phrase. Both Text 1 and Text 2 have similar items at the top of the most frequent lists, so the noun phrase summary might be something like *Anderson's film Moonrise Kingdom*.

| **Lexical Items in Text 1** | **# of lexical items** | **Chain Length (# of ties)** |
|---|---|---|
| film/MK/it/medium | 11 | 13 |
| Anderson/he | 6 | 8 |
| words | 2 | 2 |
| start | 2 | 2 |
| story | 2 | 2 |
| summer | 2 | 2 |
| world | 2 | 4 |

Table 5: Chain frequency and length for Text 1

| **Lexical Items in Text 2** | **# of lexical items** | **Chain Length (# of ties)** |
|---|---|---|
| film/MK | 11 | 18 |
| Anderson/his | 6 | 16 |
| Suzy/Hayward | 6 | 11 |
| house | 4 | 6 |
| play/played/plays | 4 | 4 |
| audience/viewers | 3 | 15 |
| scout | 3 | 10 |
| Sam | 3 | 8 |
| opens/opening | 3 | 5 |
| young love | 2 | 10 |
| camera | 2 | 7 |
| story | 2 | 5 |
| Fantastic Mr Fox | 2 | 4 |
| friend | 2 | 4 |
| Khaki | 2 | 4 |
| kid | 2 | 4 |
| rain/rainy | 2 | 4 |
| Rushmore | 2 | 3 |
| scene/sequence | 2 | 2 |
| solemn/solemnity | 2 | 1 |
| artisan/canvas | 2 | 1 |
| create | 2 | 1 |

Table 6: Chain frequency and length for Text 2

If, however, the desired summary is longer than one phrase, then additional, less frequent cohesive items can be used. In Text 1, lexical chains at the end of the text refer to the movie as a *world* that has a *summer* motif. A summary of Text 2, on the other hand, might cover many more topics, such as focusing on the two main characters, *Suzy* and *Sam* as well as characters who the various actors *play*. The *house* in the *rain* in the *opening sequence* of the film is also important in this text. Longer summaries would then be very different for the two texts.

Another type of analysis with these lexical chains can be done by mapping them to see what kinds of patterns are created. Figures 1 and 2 show a lexical chain mapping, with the location of the

lexical units shown with an "X". This analysis is a graphical representation of the chains, and it can be seen that the long, dense chains in both Text 1 and Text 2 such as *film/Moonrise Kingdom and Anderson/he* play an important role in the cohesion of both of the entire texts. However, differences are also apparent in these two texts. In Text 1, the minor lexical chains for *words*, *start* and *story*, and *summer* and *world* do not connect to each other. In Text 2, on the other hand, chains such as *Suzy*, *Sam*, *play*, and *story* act as connections between different sets of cohesive items. Even the short, dense lexical chains in Text 2 connect to each other, such as *house*, *create*, *artisanal* in sentences 1-4.

In addition, in Text 2, half of the lexical chains (11 out of 22) are represented in the final four sentences of the text, regardless of when they were first introduced. These chains include *Moonrise Kingdom*, *audience*, *Anderson*, *young love*, *Suzy*,

| | **Text 1: Sentences 1-14** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| film/MK/it/medium | X | X | X | X | X | | | X | X | | X | X | X | X |
| Anderson/he | X | X | X | | X | | | X | X | | | | | |
| words | X | X | | | | | | | | | | | | |
| start | | | | | | X | X | | | | | | | |
| story | | | | | | X | X | | | | | | | |
| summer | | | | | | | | | X | X | | | | |
| world | | | | | | | | | X | | | X | | |

Figure 1: Lexical chain map of Text 1

| | **Text 2: Sentences 1-19** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| film/MK | X | X | X | | | | | X | X | | X | | | X | X | X | | X | X |
| audience/viewers | X | | X | | | | | | | | | | | | | X | | | |
| camera | X | | | | | | | X | | | | | | | | | | | |
| house | X | X | | | | X | X | | | | | | | | | | | | |
| rain/rainy | X | | | | X | | | | | | | | | | | | | | |
| opens/opening | X | | | X | | X | | | | | | | | | | | | | |
| Anderson/his | | X | X | | X | | | | | | | | | | X | X | X | X | |
| create | | X | X | | | | | | | | | | | | | | | | |
| Artisanal/canvas | | | X | X | | | | | | | | | | | | | | | |
| scene/sequence | | | | X | | X | | | | | | | | | | | | | |
| young love | | | | | | X | | | | | | | | | | X | | | |
| Suzy/Hayward | | | | | | X | X | X | X | X | | | | | | | X | | |
| Sam | | | | | | | | | X | X | | | | | | | X | | |
| friend | | | | | | | | | X | | | | X | | | | | | |
| kid | | | | | | | | | X | | | | X | | | | | | |
| Khaki | | | | | | | | | X | | | | X | | | | | | |
| scout | | | | | | | | | X | | | | X | | | | | | X |
| play/plays/played | | | | | | | | | | X | | X | X | X | | | | | |
| story | | | | | | | | | | | | X | | | | | X | | |
| solemn/solemnity | | | | | | | | | | | | | | | X | X | | | |
| Rushmore | | | | | | | | | | | | | | | X | | | X | |
| Fantastic Mr Fox | | | | | | | | | | | | | | | X | | | | X |

Figure 2: Lexical chain map of Text 2

*Sam*, *scout*, *story, solemnity*. These terms might also be important in an extended summary of the text. By graphing the locations of the chains as was done in Figure 2, the grouping of these words at the end of the text is clear.

The fourth and final research purpose was text criticism. Whereas the first three purposes were studied using automatable methods, text criticism requires a qualitative methodology where the interpretation of the lexical cohesion in the text would be impossible using a computer. In this methodology, the most frequent cohesive items are not necessarily the focus of the analysis. Instead, the items are first organized semantically into categories such as "movie description" or "characters and actors".

For example, in Text 1, the two cohesive words in Text 1 that describe an aspect of the movie are *summer* and *world*. These two words appear in close proximity to one another at the end of the text. They paint a picture of a sunny, carefree atmosphere of "summers when kids played outside", "summer games", and "grand adventures".

In contrast, Text 2 presents a much more serious interpretation of the same movie. When *summer* is mentioned in Text 2, it is not as a reiterated cohesive item signifying playfulness, but instead as the name of the house seen in the opening credits -- *Summer's End*. As Text 2 describes the house, words such as *autumnal* and *September* are found nearby, adding to the atmosphere of changing seasons.

So while Text 1 focuses on the childlike freedom that summer brings, Text 2 instead describes the movie as the end of summer, a time of change where life becomes more serious. This can be seen in cohesive units in Text 2 such as *rain* and *solemn*. Other phrases collocated with *solemn* add to the atmosphere such as "death, abandonment" and the movie's "earnest adolescent protagonists". Through the cohesive items in Text 2, it can be seen that the protagonists are going through a change from the playful summer days of youth as they leave the comfort and protection of their families (as symbolized by the cohesive links highlighting the *house* in the *rain* in the *opening sequence*) and entering an adult world of "burgeoning sexuality" and "reckless passions".

In this way, it can be seen that the lexical cohesion of these two texts are used very differently. Text 1 leaves the reader with a positive feeling of a summertime childhood, while Text 2 is a much more serious take on the rite of passage from the fun of childhood to the somberness of adulthood.

## 4 Conclusion

This paper has discussed four different purposes for analyzing lexical cohesion in text: text evaluation, text segmentation, text summarization, and text criticism. These purposes require different methods, and each method delivers different results. For these two particular texts, two of the methods show that the lexical cohesion characteristics of the texts are the same. Some of the indices of Coh-Metrix (such as Stem Overlap of both adjacent and all sentences) give very similar results for the two texts. The Coh-Metrix results could be interpreted to show that both texts are highly cohesive compared to other similar texts. Likewise, a noun-phrase summary based on the most frequent and lengthy cohesive chains also gives the same results for Text 1 and Text 2: "Anderson's film Moonrise Kingdom".

However, all of the other methods show that the lexical cohesion characteristics of these two texts are very different. When doing a text evaluation by looking at metrics for the entire text, it was shown that Text 2 has more lexical chains. These chains are also longer, and less dense than Text 1. A moving window analysis for the purpose of text segmentation showed that the writers cover different topics in the different segments. Using lexical cohesion for text summarization gives twice as many cohesive lexical chains for Text 2 than for Text 1, meaning that a richer summary can be created for Text 2. A graphical representation of these

lexical chains also showed large differences in the ways that the lexical chains helped to tie the different parts of the text together. And finally, the qualitative interpretation of the text from the reader's perspective shows that Text 1 focuses on a happy summer motif of children's games, while Text 2 has a somber autumn feel that addresses a coming of age story.

These results point to the conclusion that although lexical cohesion appears to be a fairly straightforward concept, different purposes for using it in research can produce wildly different methods and results. This implies that lexical cohesion may not be a single construct; rather, it could comprise a cluster of several constructs, suggesting that it is a far more complex issue than it first appears. Researchers should keep these differences in mind as they decide what perspective to take when analyzing lexical cohesion in text.

## References


Barzilay, Regina & Elhadad, Michael. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, 111-121.

Graesser, Arthur C., McNamara, Danielle S., Louwerse, Max M., & Cai, Zhiqiang. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.

Hoey, Michael. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press

Hornaday, Ann (June 1, 2012). Adolescent love among eccentrics. Retrieved Nov 17, 2013 from http://www.washingtonpost.com/gog/movies/moonrise-kingdom,1221101.html

Klebanov, Beata B., Diermeier, Daniel, & Beigman, Eyal. (2008). Lexical cohesion analysis of political speech. *Political Analysis*, 16(4), 447-463.

Morris, Jane, & Hirst, Graeme. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1), 21-48.

Okumura, Manabu & Honda, Takeo. (1994). Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th conference on Computational linguistics-Volume 2* (pp. 755-761).

Luke. (October 21, 2012). Moonrise Kingdom. Retrieved September 29, 2013 from http://canetoadwarrior.blogspot.com/2012/10/moonrise-kingdom.html

Şimon, Anca, Gravier, Guillaume, & Sébillot, Pascale. (2013). Leveraging lexical cohesion and disruption for topic segmentation. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013.

Sinclair, John. (2004a). Intuition and annotation–the discussion continues. *Language and Computers*, 49(1), 39-59.

Sinclair, John. (2004b). *Trust the text: Language, corpus and discourse*. London: Routledge.

Stokes, Nicola, Carthy, Joe, & Smeaton, Alan F. (2004). SeLeCT: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1), 3-12.

Weston, Jennifer L., Crossley, Scott A., & McNamara, Danielle S. (2010). Towards a computational assessment of freewriting quality. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society (FLAIRS) conference*, 283-288.

Wong, Billy T.M., & Kit, Chunyu. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1060-1068.