A Comparison of Discourse Connective Identification of Coh-Metrix and the Penn Discourse Treebank

Ryan Callihan
Universität Tübingen
email: ryancallihan@gmail.com

Computational Linguistic Analysis of Linguistic Complexity
Core CL Hauptseminar Winter 2016
April 5, 2017

Abstract
Coh-Metrix is a linguistic tool used by many researchers to quickly measure cohesion and coherence of text. Because it is a free, easy to use, and quite efficient linguistic tool, it is widely used in academic research and analysis. The results of many of these studies are dependent on the accuracy of the Coh-Metrix tool. I will be testing the accuracy of Coh-Metrix, focusing on its analysis of discourse connectives. Discourse connectives are the easiest discourse markers to identify computationally and are a major factor for identifying the cohesion of a text. Coh-Metrix uses a "bag of words" approach to connectives, labeling every lexical connective in all possible senses. I propose that this measurement is too broad, leaving room for misinterpretation. As such, I will test Coh-Metrix with selected texts from the Wall Street Journal against the "gold standard" of discourse annotation, the Penn Discourse Treebank. Results call to question the accuracy of Coh-Metrix’ connective score.

1 Introduction

1.1 Linguistic Discourse
Discourse analysis is an often avoided topic in modern linguistics, being ignored in favor of the smaller bits of language like syntax, lexicon, or semantics. The reason for this is quite understandable. It is easier to analyze these smaller pieces because the constituents can be explicitly measured and observed within text or speech.

Discourse on the other hand is an analysis of the cohesion in a text or utterance which takes the language as a whole unit within context.
This is an area of study that has a great effect on language processing. The interpretation and meaning of text depends greatly on the previous information, genre, and context. Therefore, it is important to be able to computationally process and identify discourse markers in text.

1.2 Cohesion and Coherence

Cohesion and coherence are important topics in the conversation of linguistic discourse. Text comprehension relies heavily on a coherent mental representation found in text [Louwerse (2001)]. Coherence is the mental representation of connections. Cohesion is the lexical connective indicators found within a text which aid cohesive relationship.

According to [Louwerse (2001)], coherence cannot be based entirely upon the cohesive markers within text because it dwells in the mental representation of the reader. However, as his studies have shown, cohesion in text greatly supports coherence. Because of this fact, cohesion identification in text helps us analyze text comprehension and readability.

1.3 Connectives and Cohesion

Discourse markers can be either implicit or explicit. Implicit markers are things that cannot be identified on a lexical basis. They must be inferred from the text or utterance by the reader or listener through context and other cues.

Explicit discourse markers, on the other hand, can be identified through lexical or syntactic means. They are words or structures which connect arguments within a discourse.

[Pitler and Nenkova (2008)] found that implicit markers tend to co-occur quite often. Even though readability and coherence of text cannot rest on explicit markers alone, it stands to reason that it can have a large impact.

Discourse connectives are one of the easiest discourse markers to identify as they are a lexical measure. The more discourse connectives found in a text, the more coherent it will be, as found by [Crossley et al. (2007)]. They found that removing these explicit discourse markers in a text increased the difficulty despite being shorter.

1.4 Automatic Discourse Connective Identification

Computational readability models will therefore benefit from correctly identifying discourse connectives and their senses. One of the most used tools for calculating readability scores is Coh-Metrix (CM) [McNamara (2005)]. This will be described further in the next section. CM gives special focus to connectives due to their importance in assessing text cohesion [Graesser et al. (2004)]. The purpose of this paper is to test the CM model against the gold standard of discourse annotation, the Penn Discourse Treebank.
2 Coh-Metrix and Discourse Annotation

Developed at the University of Memphis, USA between 2002 and 2011, CM is one of the leading automatic text readability and discourse analysis tools available. Being well respected, well researched, easy, and free to use makes this program extremely popular among researchers. Searching for CM in Google Scholar yields more than 2000 results for papers about the program or citing the program. Being so trusted and used by linguists invites extra scrutiny.

2.1 What is Coh-Metrix?

CM is an automatic linguistic complexity and discourse analysis tool which analyses text cohesion on a number of different levels. The public version includes 108 indices which have been validated and theoretically grounded. Some categories include: descriptive analysis, classic readability measures, referential cohesion, latent semantic analysis, syntactic complexity, and, the focus of this paper, connectives. Users can input text of 15,000 words maximum for the public version to be analyzed within seconds. The results yield a numerical score for each index. Although these indices are backed by theory and research, most processes behind them are in a "black box" and are not open and available for public view.

2.2 Coh-Metrix and Discourse Connectives

CM has 9 connective sense indices, which include: All, Causal, Logical, Adversative and contrastive, Temporal, Extended temporal, Additive, Positive, and Negative connective incidence. These categories are based on the works of Louwerse (2001) and Halliday and Hasan (1976).

For each text, an incidence score is assigned each type of connective. The creators define the incidence score as the number of occurrences of a particular connective per 1,000 words Graesser et al. (2004). This enables us to grant an incidence score for each index.

The potential problem here is that CM factors in all lexical connectives in the score. Not all connectives in a text are discourse connectives. Furthermore, if a connective falls into multiple sense categories, it is considered as a connective within each category. For example, the connective therefore falls into the sense types: Causal, Logical, and Temporal. Therefore, it would be considered as all three sense types. This "Bag-of-Words" approach to discourse annotation could cause some problems.

3 The Penn Discourse Treebank

The Penn Discourse Treebank (PDTB) is a widely used tool in discourse studies. It will be used in this study, therefore it merits a brief introduction.
3.1 What is the PDTB?

The PDTB is a corpus of the Wall Street Journal (WSJ) consisting of more than one million words, annotated at a discourse level. All low level argument relations have been annotated, allowing researchers to do further discourse studies (Prasad et al. 2008). Because of the large amount of manual labor needed to annotate a corpus in such a way, there are not many other resources quite like the PDTB. As such, it is considered the "gold standard" of discourse annotation.

Although the PDTB is a wonderful tool, it is narrow in terms of its genre and style coverage. Most English that is produced in the world does not fit into the WSJ style guidelines. The corpus is not completely sufficient for most argument relation studies but it is enough for the purposes of this study.

3.2 Connective Annotation in the PDTB

The PDTB contains all discourse relations, implicit and explicit. Like CM, the PDTB has its own set of connective senses and many discourse connectives can have more that one sense. The top level senses are similar to those used by CM. They are: temporal, contingency, comparison, and expansion (Prasad et al. 2007).

Because of it’s thorough annotation, the PDTB will be excellent for testing the accuracy of CM discourse connective scores.
4 Testing the Effectiveness of Coh-Metrix

4.1 Method

Because CM is a closed system, there is no way to see each word or phrase that is considered a discourse connective by the system. Therefore, I identified all discourse connectives used by the PDTB and used them to measure the connective score of each file in the WSJ corpus using CM method. Then I compared the results to the actual connective annotation in the PDTB and then compared them to actual CM results.

4.2 PDTB Connectives with Coh-Metrix Metrics

The PDTB Annotation Guide does not include a list with all discourse connectives it uses. I wrote a script to pull all lexical connectives from the corpus. The PDTB includes a hypothetical connective which could be included in the implicit relations where a lexical marker is absent and I have included these as well in the list. I have included the list as an appendix to this paper in ??.

I then used the CM "bag-of-words" method to identify each potential discourse connective in each file of the PDTB irrespective of whether or not is was actually annotated as a discourse connective. The results are separated into the header sense types outlined by the PDTB, which is pictured in Figure ??.

The total number of connectives for each text were also counted. This method only counted connectives which might fall into multiple sense types once. The results using the CM method were then compared to the PDTB hand-annotated results.

4.2.1 Results: PDTB Connectives with Coh-Metrix Metrics

The difference between the "bag-of-words" method and the hand-annotated results was quite large in every category. The mean difference with the standard deviation for each sense type plus the total number of connectives can be found below.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion Count</td>
<td>31.75</td>
<td>33.36</td>
</tr>
<tr>
<td>Temporal Count</td>
<td>20.3</td>
<td>21.75</td>
</tr>
<tr>
<td>Comparison Count</td>
<td>18.5</td>
<td>19.61</td>
</tr>
<tr>
<td>Contingency Count</td>
<td>34.18</td>
<td>25.26</td>
</tr>
<tr>
<td>Total Count</td>
<td>31.83</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Table 1: Differences between CM style and hand-annotated connective occurrence
The results were then sorted in ascending order by the word count of each file to compare the connective discrepancy according to text size. Because of the relatively small number of files which are more than 1000 words, each file was assigned a number and compared the discourse connective identification in ascending order from the smallest to the largest file.

Figure 2: Connective sense difference comparison

It is clear that the CM style annotation greatly over-identified discourse connectives compared to the actual PDTB results when separated into sense types.
4.3 Actual Coh-Metrix Results and PDTB Annotation

I processed section 00 of the Penn Treebank with the CM web tool to get actual CM results. This included 99 WSJ text files but only 85 were used because the PDTB does not include every file from the PTB.

Because CM only provides a “Connective Incidence Score” per 1000 words, I reversed the formula to get the connective count. I then compared both the CM connective counts and score to the equivalent PDTB results. The sense types used by CM and the PDTB are different and do not align very well so only total connectives were identified. CM results were compared to hand-annotated and “CM style” results.

4.3.1 Results: Actual Coh-Metrix Results and PDTB Annotation

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Count</td>
<td>23.11</td>
<td>24.49</td>
</tr>
</tbody>
</table>

Table 2: Difference between actual CM and hand-annotated discourse connective occurrence

The difference between actual CM results and PDTB hand-annotated results is a bit better than PDTB connectives used with a CM method, but still not
especially accurate.

The results were also compared according to word count in ascending order.

Figure 4: CM and hand-annotated total connective occurrence comparison

Actual CM results were also compared with "CM style" results.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Count</strong></td>
<td>9.24</td>
<td>15.34</td>
</tr>
</tbody>
</table>

Table 3: Difference between actual CM and "CM style" discourse connective occurrence
The total connective counts and scores using these two methods are actually quite similar, with the PDTB granting slightly higher scores and counts.

4.4 Discussion

The results show that the CM connective identification is not an accurate metric for correctly identifying discourse connectives, especially by different sense types.

It does, however, provide a good estimate of the total connective occurrence in a text. It tends to over-estimate but it often proportionally matches the actual discourse connective occurrence pattern. Graesser and McNamara (2011) used CM to analyze texts from several genres and levels and showed that CM connective scores are a statistically significant predictor of text readability.

Actual CM results compared to PDTB CM style results produced interesting results. The results were very similar except that the PDTB in CM style over-estimated connective occurrence this time. This leads me to conclude that more words and phrases are considered to be discourse connectives in the PDTB annotation style than in the CM annotation style.

Knowing the exact connectives which CM uses would make analysis much more exact, but I believe that these results show that, although it is a useful tool, it is not an precise measure of text cohesion.

Further studies about this topic would benefit from a broader range of text genre and style than the WSJ provides.
5 Conclusion

The discourse connective identification of Coh-Metrix was compared to the gold standard hand-annotated corpus, the Penn Discourse Treebank. The results showed that Coh-Metrix is a good estimator of total connective occurrence but greatly over-identifies connectives, especially by sense type and can be used as an estimator of cohesion. It also has a finite library of discourse connectives which does not cover every possible lexical discourse marker possible.

References


Appendices

PDTB Discourse Connectives

temporalConnectivesPDTB = ['and', 'afterwards', 'eventually', 'previously', 'thereafter', 'at the same time', 'ever since', 'soon', 'as', 'when and if', 'at
the time', 'still', 'before', 'in turn', 'ultimately', 'since then', 'whereas', 'since', 'when', 'next', 'meantime', 'till', 'finally', 'now that', 'as soon as', 'if', 'until', 'much as', 'first', 'then', 'earlier', 'simultaneously', 'meanwhile', 'in fact', 'after', 'but', 'in the meantime', 'if and when', 'as long as', 'now', 'afterward', 'by contrast', 'in the end', 'at that time', 'also', 'later', 'separately', 'while', 'so', 'subsequently', 'by then', 'before and after', 'once]

expansionConnectivesPDTB = ['in addition', 'lest', 'particularly', 'indeed', 'at the same time', 'at the time', 'inasmuch as', 'still', 'yet', 'by comparison', 'except', 'meantime', 'finally', 'as it turns out', 'so far', 'for one thing', 'then', 'meanwhile', 'overall', 'on the other hand', 'in sum', 'in response', 'insofar as', 'nor', 'by contrast', 'in the end', 'in return', 'alternatively', 'as an alternative', 'separately', 'because', 'in short', 'second', 'on the whole', 'incidentally', 'further', 'ultimately', 'furthermore', 'since', 'in summary', 'simultaneously', 'however', 'besides', 'as a result', 'neither nor', 'plus', 'either or', 'consequently', 'or', 'first', 'that is', 'although', 'for instance', 'even though', 'moreover', 'similarly', 'for example', 'specifically', 'next', 'accordingly', 'therefore', 'on the contrary', 'in particular', 'until', 'on the one hand', 'but', 'else', 'as a matter of fact', 'for one', 'And', 'while', 'and', 'eventually', 'likewise', 'thus', 'as', 'if', 'in turn', 'nonetheless', 'as if', 'whereas', 'when', 'also', 'in other words', 'instead', 'much as', 'unless', 'in fact', 'after', 'additionally', 'as though', 'what's more', 'as well', 'third', 'nevertheless', 'later', 'rather', 'so', 'subsequently', 'otherwise']

comparisonConnectivesPDTB = ['and', 'because', 'regardless', 'besides', 'previously', 'on the contrary', 'thereafter', 'at the same time', 'as', 'even though', 'still', 'conversely', 'yet', 'if', 'in turn', 'on the one hand on the other hand', 'nonetheless', 'rather', 'whereas', 'meanwhile', 'when', 'meantime', 'finally', 'indeed', 'instead', 'for instance', 'by comparison', 'much as', 'then', 'earlier', 'in comparison', 'though', 'in fact', 'after', 'however', 'but', 'on the other hand', 'although', 'as though', 'nor', 'by contrast', 'in the end', 'in particular', 'nevertheless', 'similarly', 'in contrast', 'separately', 'as if', 'while', 'neither nor', 'or']

contingencyConnectivesPDTB = ['and', 'lest', 'that is', 'because', 'in short', 'indeed', 'thus', 'as', 'if', 'in turn', 'ultimately', 'for', 'for example', 'furthermore', 'since', 'when', 'finally', 'also', 'accordingly', 'now that', 'as it turns out', 'therefore', 'in other words', 'for one thing', 'plus', 'in particular', 'until', 'if then', 'then', 'when and if', 'unless', 'thereby', 'ever since', 'in fact', 'after', 'overall', 'but', 'if and when', 'besides', 'as long as', 'inasmuch as', 'in the end', 'so that', 'as a result', 'insofar as', 'as a consequence', 'for instance', 'so', 'to this end', 'eventually', 'hence', 'subsequently', 'by then', 'consequently', 'or', 'once']