Blueprints or conduits? Using an automated tool for text analysis

Stuart G. Towns
Richard Watson Todd
King Mongkut's University of Technology Thonburi

Abstract
When analyzing discourse, the researcher can take one of two perspectives: either the text of the discourse is a conduit that explicitly contains all of the meaning of the text, or the text is merely a blueprint that guides the reader in the active construction of meaning (Tomlin, Forrest, Pu, & Kim, 1997). It is more straightforward to view the text as a conduit, especially when conducting automated analyses of text which allow for the convenient investigation of large amounts of data, yet are often not able to provide much insight into how a reader might interpret the text. However, one automated tool, Coh-Metrix (Graesser & McNamara, 2011), attempts to follow the blueprint approach by taking into account reading processes and comprehension across multiple levels of language. This paper examines the application of Coh-Metrix to movie review texts to determine the usefulness of its automated analysis with respect to the blueprint model. The findings highlight some of the benefits and limitations of using automated text analysis tools.

1. Background: Automated text analysis

The past two decades has seen an explosive growth in the use of automated tools for text analysis, for example in Corpus Linguistics, as the speed and memory of computers have increased. The popularity of this field can be attributed to several factors. First, computers are able to handle large amounts of text, and are adept at sorting, counting, organizing, and searching through text very quickly. Secondly, the results of these computerized functions can be very reliable. Searching through a text for a word or counting the number of times a word appears in a text will give the same answer every time. (This assumes that the same search or counting algorithm is used every time. The problems encountered when using different algorithms will be discussed later in this paper.) This high level of reliability of computerized algorithms also reduces the chance of adding human error or bias to the results. Finally, corpora can be annotated in a computer-readable way, such as by adding part of speech tags to a text, to allow for more complex analyses (McEnery, Xiao & Tono, 2006).

There are many automated text analyses that are currently available to the linguistics researcher. Going from very simple analyses to more complex ones, there are word lists, frequency lists, keyword analysis, word cluster analysis, concordances, keywords in context (KWIC), and collocation searches (Bowker & Pearson, 2002). A researcher can also use part-of-speech taggers, find similar and related words using WordNet, or compute the lexical or semantic complexity of a text. The tools that can help automate these tasks can be in the form of software that can run on your computer, such as AntConc (Anthony, 2011) and Wordsmith Tools (Scott, 2012), or they can be web-based, such as the search and concordancing functions of the Corpus of Contemporary American English (Davies, 2008).

1.1 Limitations in using automated tools to analyze text

Despite the advantages of using automated analysis tools, there are many limitations and criticisms against them as well. One major limitation is that today's computers are not yet able to fully process and understand the meaning in text, especially with respect to the socio-cultural context in which the text was written.

This distinction can be seen from the perspective of a discourse interaction model developed by Tomlin, Forrest, Pu, & Kim (1997), known as the conduit and blueprint discourse metaphors. In the conduit metaphor, the producer of a text has placed all intended meaning in the text itself. The text is merely a pipe that carries all necessary information to the audience, who is then able to fully understand the meaning from the words themselves without any outside knowledge about the author, or the world, or the social context surrounding the text. Therefore, in order to analyze a text that is viewed as a conduit, the text itself is the only possible source of data.

In contrast, the blueprint model of discourse interaction sees the text as a guide which aids the audience in creating the meaning of the text for themselves. With the help of the text, the audience can then construct the meaning of the text by using building materials such as their past experiences and world knowledge (the task of "knowledge integration") as well as by using the cues in the text which direct them as to how to organize the information into a coherent whole (the task of "information management").

Tomlin et al. consider the conduit method to be a "naive" way to view discourse. Flowerdew (2005) echoed this criticism, saying that using automated tools to analyze text (for example, using concordance software) strips away the social context from the text in the concordance lines. This greatly hinders the investigation of the full meaning of the text, which is often found in the socio-cultural context surrounding it. Flowerdew went on to say that another limitation is that concordance lines creates a "somewhat atomized, bottom-up type of investigation". This is in stark contrast to, say, a genre analysis which starts with larger sections of the text as an investigative unit in order to consider the socio-cultural context (the discourse community) in which the producer and the audience reside.

1.2 An automated text analysis tool: Coh-Metrix

There are many different automated tools that are available to the researcher such as word frequency counters, keyword analyzers, concordances, collocation searches, and lexical and semantic complexity analyzers. But perhaps the one tool with the widest range of functionality is Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004; Graesser & McNamara, 2011), which claims to take cognitive and social constructs into consideration when analyzing a text and thus possibly follows a blueprint model of discourse. The original purpose of Coh-Metrix was three-fold: 1) to offer many different automated analyses using several different knowledge sources in one tool, 2) to update the idea of "readability" to consider more modern theories on text and discourse rather than only relying on surface features such as word and sentence length, and 3) to be the first automated tool to measure text cohesion (McNamara & Graesser, 2010).

The current version of Coh-Metrix (Version 3.0) is available online for any researcher to use. After the researcher submits a text that is shorter than 15,000 characters, the Coh-Metrix website will analyze the text and will return a single number for each of 108 indices located at several different levels of language. The developers of Coh-Metrix have organized these 108 indices into 11 groups, as seen in Table 1.

Table 1 The 11 categories determined by the Coh-Metrix developers along with example indices for each category (Not all of the 108 indices are shown here.)

Coh-Metrix Output Category	*Example Coh-Matrix Indices for each category*
Descriptive	Word, sentence, and paragraph count and length
Text Easability Principal Component Scores	Text Easability PC Syntactic simplicity, Text Easability PC Deep cohesion
Referential Cohesion	Noun overlap, argument overlap, anaphor overlap
Lexical Semantic Analysis	LSA overlap in adjacent sentences, LSA given/new
Lexical Diversity	Type-token ratio for content words and for all words
Connectives	Causal, logical, temporal, and additive connectives
Situation Model	Causal and intentional verbs, WordNet overlap
Syntactic Complexity	Left-embeddedness, number of modifiers per noun phrase
Syntactic Pattern Density	Noun, verb, adverbial, prepositional phrase densities
Word Information	Age of acquisition, familiarity, concreteness, polysemy, hypernymy
Readability	Flesch Reading Ease, Flesch-Kincaid Reading Level

Not only does Coh-Metrix provide a wide range of analyses at several different levels of language, it can also be very efficient and effective. With regards to efficiency, Coh-Metrix is able to analyze up to 15,000 characters at one time, with the total analysis taking three to five minutes. Even doing one of the 108 indices by hand probably would take longer, let alone doing all 108. So there is no doubt that using automated tools such as Coh-Metrix is a very efficient way to investigate large amounts of text.

Coh-Metrix can also be considered more effective than a manual analysis as it uses many well-known outside sources to tag and annotate words in the text. The Brill (1995) POS tagger is used to identify parts of speech for every word in the text. WordNet (Fellbaum, 1998) is used to identify causal verbs as well as polysemy and hypernymy of the words. And the MRC Psycholinguistic database (Coltheart, 1981) supplies information to Coh-Metrix about the familiarity, concreteness, imageability, and meaningfulness of words. The results from these and other outside sources could possibly be replicated manually by the researcher, but using a well-tested database that has been developed over many years gives more reliability to the results.

Although Coh-Metrix is an excellent choice of an automated tool to use for text analysis, there are still limitations that the researcher should consider. If we view the conduit vs blueprint models of discourse interaction (Tomlin, et al., 1997) as a continuum rather than a dichotomy, we can see that various Coh-Metrix analyses fall on different places along the continuum. At the extreme end of the conduit perspective where only the text is considered and no outside sources are used, we find Coh-Metrix indices such as word count and type/token ratio. These two analyses take a conduit approach because a researcher or an automated tool does not need any world knowledge or any knowledge of social constructs or contexts to be able to count how many words are in a text. Only the text itself is needed. But as we start to bring in other data to help us analyze the text, we move towards the other end of the spectrum towards the extreme blueprint model. Using WordNet to tag similar words, or using the MRC Psycholinguistic database to determine imageability is starting to consider knowledge outside of the text. But as can be seen in the diagram below, we never quite reach the full blueprint model, due to the primitive linguistic processing capabilities of today's computers.

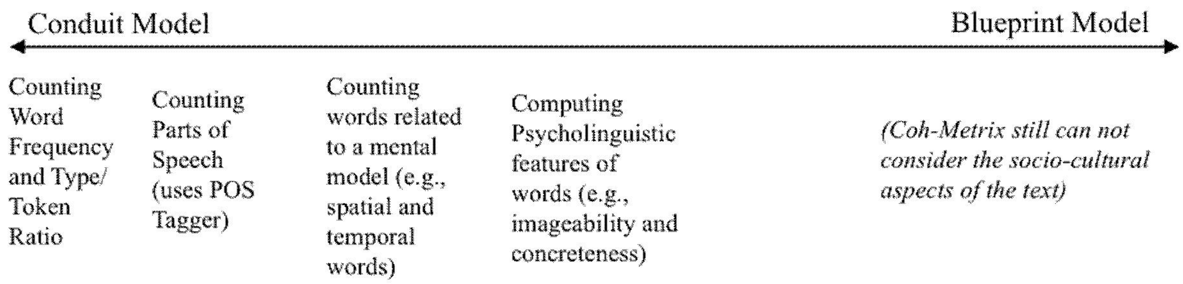

Figure 1 The continuum between a conduit and a blueprint model of discourse interaction with sample Coh-Metrix analyses

Even though some Coh-Metrix analyses attempt to do more than just view text as a conduit, there are still limitations that must be taken into consideration by a researcher using Coh-Matrix. The purpose of this paper is to share some issues that we have uncovered. In the following sections of this paper, we will discuss a methodology for using an automated text analysis tool. We will then discuss the use of Coh-Metrix in our investigation into the writing quality of Hollywood movie reviews. The paper will then conclude with a discussion of how human intuition plays an important role in all three steps in our methodology.

2. Methodology for using an automated tool such as Coh-Metrix

Our research interest is to investigate writing quality. To do this, we created two corpora: one with Hollywood movie reviews that were deemed "higher quality" because they were written by Pulitzer Prize winners and another with movie reviews of "lower quality" which were taken from movie review blogs. (All movie reviews were publically available on the Internet.) Since we wanted to analyze a large amount of text quickly, we first started with automated tools such as Coh-Metrix. Our methodology for using automated text analysis tools consists of three steps, as seen in Figure 2.

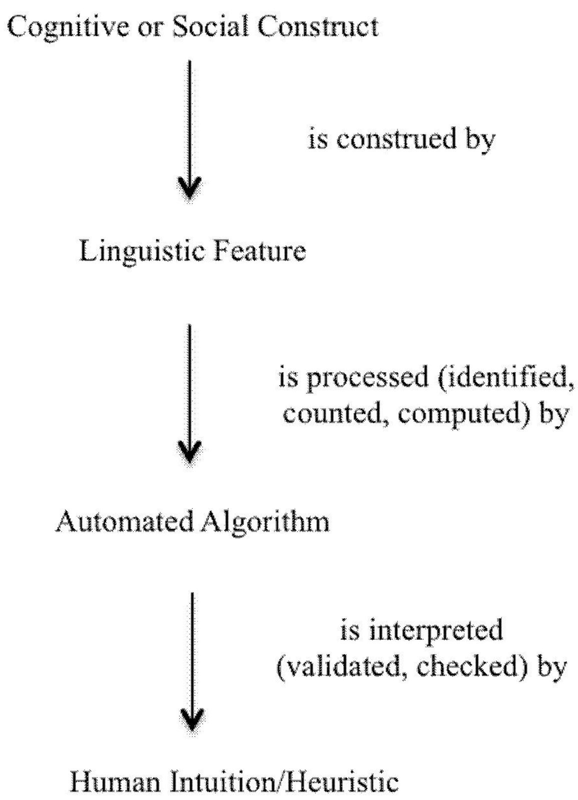

Figure 2 A three-step methodology for the use of automated text analysis tools.

2.1 Step 1: Cognitive or social constructs are construed by linguistic features

One of the original purposes for the creation of Coh-Metrix was to improve the readability indices that were available at the time. Therefore, many of the Coh-Metrix indices are founded on a cognitive construct (such as comprehension) that is affected by linguistic features of a text. For example, the more words that are found in a sentence before the main verb, the higher the cognitive load is for the reader. That is, left-embedded sentences cause the reader to store a large amount of information in memory before getting to the main clause of the sentence. Other features that increase cognitive load are word length, passive voice, logic words, and dissimilar sentence structures (Graesser & McNamara, 2011).

We started our research with the assumption that writing quality (that is, the interpretation of writing quality as judged by the reader) is affected by linguistic features that are found in the text. These linguistic features could be at any level of the text, from word choice on the lexical level to the use of cohesive devices to guide the reader on the discourse level. Coh-Metrix seemed like an ideal automated text analysis tool to differentiate which linguistic features are used differently between the movie reviews written by Pulitzer Prize-winning authors and those written by bloggers.

2.2 Step 2: Linguistic features are processed by an automated algorithm

The Coh-Metrix tool is publicly available on an Internet website. To use it, the researcher simply enters the text into a text box on the website, and clicks a button to submit the text to the

Coh-Metrix server. Coh-Metrix then uses many different automated algorithms in order to deliver a single number for each of 108 indices. However, researchers using the website only see their input and Coh-Metrix's final numerical output. They do not have access to the hidden automated algorithms. The algorithms can be simple (e.g., counting the number of words in a text) or complex (e.g., lexical semantic analysis, which involves vector math). The algorithms can also vary on the use of outside knowledge sources such as POS taggers, WordNet and psycholinguistic databases, or not use any at all. But again, the use of these outside knowledge sources is hidden from the user. For example, even though every word is tagged with a part of speech, the researcher never sees these tags.

The Coh-Metrix website restricts the text input to a maximum of 15,000 characters. Our movie reviews were, on average, a little less than 1,000 words or around 4,000 characters each. Therefore, since grouping the reviews by corpus, by author, or by movie would have been too large, we analyzed each movie review separately. These reviews came from 5 Pulitzer Prize-wining authors and 5 bloggers, and covered the same 12 recent Hollywood movies. The processing of each movie review took around three to five minutes to complete and the results were returned in a large table with one number given for each of the 108 indices.

2.3 Step 3: Algorithm results are interpreted by the researcher

As noted in Step 2 above, the results from Coh-Metrix are purely quantitative, which is common for most automated corpus linguistics tools. However, due to the limitations discussed earlier, many researchers have pointed out that there must also be a qualitative step in the research process where the researcher analyzes the results using human intuition (Baker, 2006).

This is certainly true for our research as well. As we were looking for differences between the "higher quality" and "lower quality" texts, our preliminary analysis has identified 18 indices which may have significant differences between the two corpora. Most of these indices were in the Referential Cohesion category, so we have flagged this area to look at in more detail in a future qualitative analysis.

3. Comparing movie review corpora with Coh-Metrix

As described in the previous section, we used the three-step methodology to investigate the linguistic differences between "higher quality" Hollywood movie reviews written by Pulitzer Prize winning authors and "lower quality" movie reviews written by Internet bloggers. In order to do this, we collected a total of 101 movie reviews (49 "higher quality" and 52 "lower quality") on the same 12 Hollywood movies. We processed each review through Coh-Metrix separately. We then computed a t-test comparing the Pulitzer Prize wining reviews to the blogged reviews. Out of the 108 indices computed by Coh-Metrix, 18 of them had a p-value of less than .0000001. These 18 indices and their category from Table 1 can be found in Table 2.

As can be seen from Table 2, the category of Referential Cohesion had the largest number of indices with $p<.0000001$, while Lexical Diversity, Connectives, Situation Model, and Syntactic Pattern Density had none.

The indices which can be considered taking a conduit approach to discourse such as word count and sentence length are straightforward to interpret. In our data, the average word count between the higher quality and lower quality corpora was not significantly different, with an average of 975 and 893 respectively. One conduit-based index that did show a significant difference was sentence length, with an average of 18 and 22 respectively.

Table 2 Coh-Metrix indices with p<.0000001 comparing "higher quality" to "lower quality" movie reviews

Coh-Metrix Output Category	*Example Coh-Matrix Indices for each category*
Descriptive	Sentence length
Text Easability Principal Component Scores	Text Easability PC Syntactic referential cohesion z score Text Easability PC referential cohesion, percentile
Referential Cohesion	Noun overlap adjacent sentences, binary Noun overlap all sentences, binary Argument overlap adjacent sentences, binary Argument overlap all sentences, binary Stem overlap adjacent sentences, binary Stem overlap all sentences, binary Content word overlap adjacent sentences, proportional Content word overlap all sentences, proportional
Lexical Semantic Analysis	LSA overlap all sentences in paragraph
Lexical Diversity	None
Connectives	None
Situation Model	None
Syntactic Complexity	Minimal Edit Distance, all words Minimal Edit Distance, lemmas
Syntactic Pattern Density	None
Word Information	CELEX [log] word frequency for content words Polysemy for content words Hypernymy for content words
Readability	Coh-Metrix L2 Readability

Other indices are more difficult to interpret. Many of them, such as those for Referential Cohesion, are measured on a scale of 0 to 1. On average, the higher quality blogs had lower scores for cohesion than the blogs. But without analyzing the texts manually using human intuition, it is difficult to understand the meaning of the numbers. Therefore, the third step in our methodology is very important. For our research purposes, the second step of using automated text analysis tools merely gives us some pointers as to where to focus our qualitative investigations. We also realized that human intuition is important at every step, as the next section will show.

4. The need for human intuition

In the three-step methodology we employed for this study, the final step whereby the researcher must use human intuition to analyze the results given by automated tools is an "essential step" (Biber, 1998). However, based on our results, we realized that human intuition is not only crucial at the end of the process, but is important at every step. Every point in the research should include a careful analysis guided by the researcher's intuition, as described below.

4.1 Human intuition at step 1: Construct validity

As with any kind of research, the researcher must be careful to select an appropriate tool or instrument to collect and analyze data. The researcher must use human intuition to insure that the tool or instrument is aligned properly with the research purpose, which is grounded in a specific cognitive or social construct. There is a danger that the ease of analysis using automated tools and the quantitative nature of the results will cause the researcher to easily accept the validity of the results without critical thought.

For example, for many years, it was believed that there was a strong relationship between word and sentence length and the readability of a text. Readability formulas such as the Flesch-Kincaid Grade Level were completely based on word and sentence length (Klare, 1974-1975, as cited in Graesser & McNamara, 2011). However, intuitively we now know that many other factors are involved in determining the actual readability of a text, including both other linguistic features as well as the reader's own background and knowledge. Coh-Metrix still, however, includes descriptive indices such as word, sentence, and paragraph length, and readability calculations. These analyses are purely based on the conduit perspective of discourse as no knowledge outside of the text is needed to perform the analysis. Therefore, it is up to the researcher's intuition on whether or not these quantitative measures are truly applicable and valid for their research purpose. If the research is viewing the text as a blueprint, then the conduit-based perspective of word and sentence lengths should probably not be the only focus of the study.

4.2 Human intuition at step 2: Potential problems with the algorithms

Coh-Metrix, like most automated text analysis tools, does not give the researcher access to or allow the researcher to view the actual computations and processing. Researchers must simply have faith that the results are correct. There are three potential areas where this can be an issue: 1) it is not clear as to what the algorithm is actually doing and therefore it might be doing something unexpected, 2) the algorithm might be incorrect or have bugs which would cause the wrong answer to be given, and 3) the algorithm might not be able to be completely accurate due to technological limitations. Each of these three cases requires that the researcher to use human intuition to insure the reliability and validity of their results. We will look at each case below.

4.2.1 Lack of clarity of what the algorithm is doing

Naturally occurring language is so complex and varied that there are many ways to determine what exactly a word or a sentence is. Because of this, before a text is analyzed, it should go through a non-trivial pre-processing stage to normalize the text (Grefenstette & Tapanainen, 1994). In the case of Coh-Metrix, the processing is hidden from the researcher, so it is not clear what pre-processing (if any) has been done on the text before the analysis starts. And as far as we know, the algorithms that Coh-Metrix uses to determine what is counted as a word or a sentence have not been publicly documented. Only by testing one word at a time in Coh-Metrix can we see that the system counts "don't" as two words and "birds-eye" as one. This is not to say that these algorithms are wrong, it's just to say that researchers should be aware of how these units are defined by whatever tool they are using as it may affect the results of their study.

Likewise, the analyses that are closer to the blueprint model of discourse use outside knowledge such as parts of speech tags or information from psycholinguistic databases, but the researcher does not know exactly how this knowledge is being used in the analysis. There is no easy way for the researcher to check whether or not the outside knowledge is being used correctly, or how it might affect the research results or purposes.

4.2.2. Incorrect results

With a set of results of 108 numbers from every Coh-Metrix analysis, it is tempting to trust that all of the numbers are actually correct. In our study, one of the indices that showed a significant difference between the Pulitzer Prize winner reviews and the blog reviews was sentence length. To double-check the results, we ran a small test of one long sentence which we could check by counting the number of words per sentence and compute the average sentence length by hand. The text we chose was the opening paragraph of Ann Hornaday's review of The Avenger's movie in the Washington Post.

> "There were ripples of anticipation — and some anxiety — when Marvel Enterprises announced that Joss Whedon would direct "Marvel's The Avengers," the comic-book-movie to end all comic-book-movies, featuring Iron Man, Captain America, the Hulk, Black Widow, Hawkeye and (did I miss anyone?), oh yes, Thor." (Hornaday, 2013).

It is easy to see that this is one long, complex sentence made up of either 44 or 48 words, depending on whether or not the algorithm counts "comic-book-movie" as one word or three. (This is another example of the issue of lack of clarity discussed in the previous section.) Coh-Metrix, however, surprisingly analyzed this as being four sentences with 45 total words and an average sentence length of 11.25 words per sentence. This problem seems to be deeper than just the operationalization of the meaning of a word or a sentence.

To get a better understanding of the validity of these results, we conducted a trial-and-error analysis with this example sentence. We were able to determine that the possessive noun "Marvel's" was counted as two words, which caused our word count to be off by one from what we would expect. For the sentence breaks, we determined that both dashes and question marks were counted as sentence breaks. This one sentence had two dashes and one question mark, therefore it was counted as four sentences. The addition of a question mark as a sentence break in this case was especially troubling, as it meant that Coh-Metrix analyzed the question as a sentence inside of another sentence.

We were surprised that both a possessive form of a noun was counted as two words, and a dash was considered a sentence break. Since several of Coh-Metrix's indices use number of words or number of sentences as a variable in the calculation of other indices (including many of the indices that showed significant differences between corpora in our study), these results are very concerning.

4.2.3 Technological limitations

Unfortunately, the state-of-the-art in automated text analysis is still not very reliable when it comes to some of the indices that Coh-Matrix attempts to process. With respect to the conduit-blueprint continuum, the closer an analysis is to a full blueprint model, the less reliable an automated algorithm will be. One good example of this is the Coh-Metrix index called Anaphor Overlap. This index tries to match a pronoun in one sentence with the noun that it refers to in the previous sentence. Human intuition is able to solve this problem, often without any trouble, but there are no heuristics or automated algorithms that can give perfectly reliable results. To their credit, the developers of Coh-Metrix have briefly mentioned this limitation in Anaphor Overlap (Graesser & McNamara, 2011), but an issue like this is not necessarily obvious to someone using the tool. To avoid this problem, researchers should improve their understanding and intuition about the limits of technology for language processing.

4.3 Human intuition at step 3: Results interpreted qualitatively

Most Applied Linguistics research is grounded in a social context and is therefore viewing discourse from the perspective of the blueprint model. However, as the previous sections have shown, automated analyses are very limited in their ability to process text viewed in this way. To solve this problem, many researchers have pointed out the need to have a qualitative analysis step after using automated text analysis tools. This is true for Coh-Metrix results as well. Every Coh-Metrix analysis is made up of 108 numbers that represent 108 different indexes. These numbers only have a relative meaning to other numbers (e.g., being higher or lower, or significantly different from one another) but since the interplay between these 108 indices is so complex, the numbers themselves do not provide much insight into the text without the researcher taking into account the socio-cultural context in which the text was written.

5. Conclusion

Recent advances in computer technology have given researchers many new techniques and tools for automated text analysis. Researchers can now easily search through millions of words of authentic natural language, or create frequency lists or keyword lists, or organize a text into concordance lines in a matter of seconds. However, these tools and techniques still view text from the naive perspective of discourse interaction as a conduit. Tools such as Coh-Metrix can go a few steps towards a blueprint model by bringing in multiple outside sources such as WordNet and psycholinguistic databases as a proxy for the real world knowledge that every speaker and audience have, but the tools are unable to reach the extreme blueprint end of the spectrum.

Researchers should be aware of these limitations as they include automated text analysis tools in their methodologies. These methodologies can take the form of a three-step process of: 1) cognitive or social constructs are construed by linguistic features, 2) linguistic features are processed by automated algorithms, and 3) the results from the automated algorithms are analyzed qualitatively by human intuition. Researchers should be aware, however, that their human intuition is an important feature of each of the three stages to insure that the methodology and tools chosen match the research purpose as well as give valid and reliable results.

Acknowledgements

Special thanks to Woravut Jaroongkhongdach, Dougal Graham, and Daron Loo for very helpful suggestions on previous versions of this paper.

References

Anthony, L. (2011). AntConc (Version 3.2.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Biber, D. (1998). *Variation in Speech and Writing*. Cambridge: Cambridge University Press.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. London: Routledge.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21, 543–565.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4), 497-505.

Davies, M. (2008) The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24(3), 321-332.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371-398.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193–202.

Grefenstette, G., & Tapanainen, P. (1994). What is a Word, what is a Sentence?: Problems of Tokenisation (pp. 79-87). Rank Xerox Research Centre.

Hornaday, Ann (May 4, 2012). Superheros with team spirit. Retrieved November 17, 2013 from http://www.washingtonpost.com/gog/movies/marvels-the-avengers,1059548.html

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based language studies*. New York: Routledge.

McNamara, D. S., & Graesser, A. C. (2010). Coh-Metrix: An automated tool for theoretical and applied natural language processing. *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*. Hershey, PA: IGI Global.

Scott, M., 2012, WordSmith Tools (Version 6) [Computer Software]. Liverpool: Lexical Analysis Software. Available from http://www.lexically.net/wordsmith/.

Tomlin, R. S., Forrest, L., Pu, M. M., & Kim, M. H. (1997). Discourse semantics. In T. A. van Dijk (Ed.), *Discourse as structure and process (Vol. 1)*. Sage.