Methods for identifying relations and associations in text

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

Abstract

For decades, linguists have conducted research into the connectedness between concepts in a text. Much of this work has analyzed generic logical relations such as those outlined by Cruse (2011) or lexical cohesion relations described by Halliday and Hasan (1976). However, in order for content to be developed through discourse, the words in a text must also have conceptual associations with one another in addition to these generic logical relations. To date, there has been very little research done on conceptual associations, perhaps in part because they are difficult and time-consuming to identify. To aid in this identification, this study investigates the use of several different computer-aided resources to analyze the relations and the associations between words in a short text. The automated resources include: a digital thesaurus, WordNet, semantic tagging using the UCREL Semantic Analysis System, a word association database, near neighbors scores using Latent Semantic Analysis, and MI scores from COCA. It was found that all six methods gave results that represented the connectedness in the text, with the thesaurus being the most valid of the six. The computer-aided results also showed high corroboration with results from similar manual analyses which were based on researcher intuition.

1. Background: Proficient vs Exceptional Writing

This paper is part of a larger study into the differences between proficient and exceptional writing. The texts which were analyzed in the larger study were movie reviews, with proficient writing being movie reviews written by bloggers while exceptional writing was represented by movie reviews of the same movies written by Pulitzer Prize winners. Multiple linguistic features across syntax, lexis, and discourse were compared between these two corpora of proficient and exceptional texts in an attempt to uncover differences between the two. One of the linguistic features which seems to show large differences between the two corpora of movie reviews is the connectedness of the concepts in the texts, both in the frequency of connectedness throughout the text as well as the types of connectedness used, with the exceptional texts showing much more frequent and much richer use of connected concepts.

The analysis of the connectedness of concepts in the texts in the larger study was done manually by analyzing each text by hand and highlighting connected concepts based on the researcher's intuition. This analysis was a time-consuming process which greatly limited the number of texts which could be investigated. However, there are several available lexical databases that contain information about different kinds of connectedness, and using these computer-based sources to analyze the connectedness in texts could potentially allow for a larger amount of data to be analyzed within a shorter timeframe. Therefore, this paper will investigate the use of multiple automatable methods to analyze a segment of one of the exceptional movie reviews from the larger study. The automatable methods will be judged as to which source most reflects the organization of the lexis in the selected text. The combined results from all methods will also be compared to the original manual analysis of the connectedness in the text to judge the validity of using these methods to identify connectedness.

2. Cohesion and Coherence: Reiterations, Relations, and Associations

In the field of linguistics, the connectedness of concepts in a text is traditionally viewed from the perspective of cohesion and coherence. Cohesion relates to the surface features of a text, that is, the way in which concepts are connected to each other in the text, while coherence is related to the connection of concepts in the reader's mind. Since both cohesion and coherence are concerned with the connection of concepts, they can be viewed as a continuum going from the cohesion of explicit connected concepts in the text to the coherence of the implicit meanings of the text in the mind of the reader (Watson Todd, 2016). At the extreme case of cohesion, no background knowledge is needed on the part of the reader to understand the connections between concepts, while the extreme case of coherence relies on large amounts of background knowledge to draw implications of meaning from the text.

Halliday and Hasan (1976) defined five types of cohesion: lexical, reference, substitution, ellipsis, and conjunction. Of these five types of cohesion, this study (and the larger study above) is concerned mostly with the lexical cohesion found in texts. Lexical cohesion is an often studied phenomenon with many different applications, including investigating writing quality by finding the correlation between the cohesion in a text and human evaluations of the text (Weston, Crossley & McNamara, 2010), NLP applications such as topic segmentation, which look for breaks in the lexical cohesion in a text to signify the change of topic in the text (Şimon, Gravier & Sébillot, 2013) and text summarization, which is aided by lexical cohesion in identifying the important topics in a text. Lexical cohesion can be used in qualitative analyses as well, such as a method for identifying metaphors in political speeches (Klebanov, Diermeier & Beigman, 2008).

The most extreme case of lexical cohesion in a text (i.e., the most explicit connection between two words) can be found in the exact repetition of words in the text, assuming that the repeated words are not polysemous in that context. In this extreme case of connections through repeated words, the reader does not need any world knowledge or background knowledge to know that the repeated words represent the same concept. This study will refer to this type of connection as a reiteration. Another way to reiterate a concept in a text is to refer to it using other words, such as a pronoun referring to a previous noun. Connections in a text can also be created using classical lexical relations. In this case, relations refer to connections which can be linked by logic such as synonyms which could be seen as equivalence, hyponyms which are often linked by entailment or meronyms which are a relation of inclusion (Cruse, 2011). These types of relations might go by other names, such as paraphrases and semantic associations (Hoey, 2005) but the defining feature of relations is that they are logical connections. And although these connections are considered as lexical cohesion along with reiteration connections above, they require more world knowledge on the part of a reader than simple reiterations. In addition, there may be even further connections between concepts in a text other than reiterations and relations. These connections between different concepts exist in the mind of the reader rather than in the text itself. This study will refer to these connections as associations, following the nomenclature of Watson Todd (2013).

So in this way, we have three main categories of connectedness in a text: reiterations, relations, and associations. On the cohesion side of the continuum, reiterations require the least amount of world knowledge as they come from the surface features of the text such as between repeated or referring words, while on the coherence side of the continuum, associations require the most amounts of world knowledge since they are connections created in the mind of the reader. Out of these three types of connections, identifying reiteration such as repeated words or pronouns which refer to previous nouns in a text in a text is fairly straightforward and does not provide much understanding of how concepts are linked in a text. Therefore, in order to gain a broader understanding of how topics are developed in a text, this paper will not attempt to identify reiterations, but instead will focus on uncovering the connections made by relations

and associations. This will be accomplished by using several different lexical resources which all have the potential for being automated, allowing for large amounts of texts to be quickly and easily analyzed. Some methods will primarily be used to identify relations, while others will be primarily used to identify associations.

To identify word relations, lexical resources such as thesauruses and lexical databases such as WordNet can be used. These resources are hierarchical representations of the lexis in a language determined by the relations between words. Thesauruses focus on synonyms and antonyms, while WordNet takes a broader view by also including other relations such as hyponyms/hypernyms, meronyms, and troponyms. Another potential source of word relations might be found by using a semantic tagger to group words semantically. Inside each semantic group, logical relations such as synonyms/antonyms or hyponyms/hypernyms might be found. However, the hierarchy used by a semantic tagger will also consider other types of semantic connections, and therefore some word associations might also be found inside specific semantic groups. In this way, a semantic tagger can help to identify both relations as well as associations.

Identifying word associations poses a more difficult problem since they are based on the connections that are made in the reader's mind, rather than being surface features such as reiterations, or based on logical relations such as synonyms. There are, however, various ways that we might be able to uncover potential word associations in a text. Research into word association have been conducted wherein participants are given a word and asked to respond with an associated word (Nelson, McEvoy & Dennis, 2000). These word association pairs can then be stored in a database and searched. Another way to uncover word associations is to use Latent Semantic Analysis (LSA) to uncover words which tend to appear in the same text as another word. A high near neighbor score in an LSA analysis might show that these words are conceptually associated (Landauer, Foltz, & Laham, 1998). A third method is to use a large-scale corpus to find words which appear with a greater than random frequency in reasonable proximity. By searching for high Mutual Information (MI) scores of word pairs in a text, these associations can be found.

The remainder of this paper will discuss the current study which attempts to uncover word relations and word associations in a short sample text. Specifically, word relations will primarily be investigated using the Oxford American Writer's Thesaurus (Lindberg, 2004), WordNet (Princeton, 2010), and the UCREL Semantic Analysis System (UCREL, n.d.), while word associations will primarily be investigated using the word association database at Small World of Words (Small, n.d.), the Near Neighbors LSA tool hosted at the University of Colorado (Latent, n.d.), and MI scores from the Corpus of Contemporary American English (COCA) (Corpus, n.d.). We will then conduct an evaluation of the results from these six different sources, with the goal being to determine the most valid method(s) for a specific text as well as an overall evaluation of how well these sources were able to identify the connectedness in the text, based on our researcher intuition about the relations and associations present in the text.

3. Text selection and analysis criteria

As mentioned earlier, this study is a part of a larger study which is investigating the differences between proficient writing and exceptional writing. It was found that, in general, exceptional writing contains more relations and associations than proficient writing, which relies mostly on reiteration. Therefore, in order to investigate the use of automated methods to uncover relations and associations, a short segment of an exceptional text which was found to have a high amount of connectedness was used for this current study. The text chosen for analysis is the opening 218 words of a review of the movie Moonrise Kingdom directed by Wes Anderson. The review was written by Pulitzer Prize winner Ann Hornaday and was published in the Washington Post newspaper. The text to be analyzed is as follows:

> "Moonrise Kingdom" opens with no music — just the sound of raindrops falling on the roof of a preternaturally cozy house, which the camera gently leads the audience through as the family members inside go about their rainy day business.
>
> Bathed in apple reds, egg yolk yellows and an air of studied eccentricity, the house is immediately recognizable as yet another habitat created by Wes Anderson, a film director whose obsession with material culture, nostalgia and nursery comforts borders on the fetishistic.
>
> Of course, for viewers who happen to share Anderson's taste for boldly framed, bespoke productions — in which everything looks (and most probably is) lovingly handmade and artisanal, "Moonrise Kingdom" will simply offer yet another chance to live, at least for a little while, in the kind of universe only Anderson can create. (You can almost smell the damp canvas and wood polish in that opening sequence.) Those who long ago wrote off the writer-director as insufferably mannered and arcane — the usual term of art is "twee" — well, they're welcome to stay out in the rain.
>
> That opening scene house has a name, by the way: Summer's End, which turns out to aptly capture a vaguely autumnal tale of young love that takes place in early September 1965 — a time of Ford Falcons and mothers who smoked (Hornaday, 2012).

The first step of the analysis is to decide which words in the text we will attempt to match through relations or associations. The focus of the analysis is the concepts in the text, so it makes sense to start with content words (nouns, verbs, adjectives, and adverbs) and to ignore function words. However, not all content words were analyzed -- adverbs, proper nouns, auxiliary verbs, phrasal verbs, or other idioms (e.g. *of course*, *by the way*) were not included in the analysis. In order to further facilitate the analysis, the words chosen were required to be separate entries in reference sources such as dictionaries and thesauruses. Therefore, hyphenated nouns phrases acting as adjectives (e.g. *rainy-day*) were broken up into separate words, unless they expressed a single concept that was found in a dictionary (e.g. *egg-yolk*). Words that are found in the top 250 most common English words were also not analyzed, as these words tend to be the most polysemous, and multiple meanings will tend to give too many false positives when identifying relations and associations. By following these guidelines, a 73-word list taken from the above text was created in order to perform further analysis with our six automated lexical resources.

As mentioned earlier, all of the methods and sources used for the six lexical resources are automatable since they are either look-ups in a database (e.g., in the thesaurus, WordNet, and association experiments) or computed using various algorithms (e.g., using semantic tagging, LSA, and MI). For this study, however, only partial automation was used, such as gathering data by doing manual searches through a website interface, and then manually performing the analysis in a spreadsheet. For future studies, both the database look-ups and the analysis methodologies have the potential for full automation that could quickly return full results based on any inputted text.

4. Results

This study contained three main phases. Once the word list was created, the first main phase was to search for each word using each of the six sources, copy all potential connections to a spreadsheet, then search all connections for words which are in the word list, giving us in-text word pairs. Four out of the six sources returned a limited set of potential word connections,

and the numbers of in-text word pairs were as follows for these four sources: 21 pairs were found in the thesaurus, 32 pairs were found in WordNet, 21 pairs were found with semantic tagging and 45 pairs were found from the associations database. The other two sources of LSA and MI scores, however, are both based on the results of searching a large corpus for words which appear near each other, and therefore the number of resulting word pairs could be in the thousands, depending on how different search variables are set. So in order to get a similar number of results to the other four sources, quantitative cut-offs were used. For LSA, only connections with a near neighbor score of greater than or equal to 3.5 were used, resulting in 34 word-pair matches. And for MI scores, only words with a frequency in the corpus of at least 10 and MI scores greater or equal to 3.0 were considered, which returned 28 word-pair matches.

However, not all of the connected word pairs that were identified from these six sources are relevant relations or associations in this specific text. An example of this are the two connected word pairs of *audience-house* and *audience-chance*. In the first case, the definition of *house* in this connection is an *audience* at a theatre. But in our text, the *house* is the building where a family lives. Likewise, for the connection of *audience-chance*, an *audience* is the *chance* to have your case heard in court, but in our text, the *audience* are those who are watching the movie. So while *audience-house* and *audience-chance* are hyponymic relations (according to WordNet), they are not considered to be connections in the context of our text.

So the next step is to determine which of the identified word pairs are relevant to our text. This is a manual process that requires the intuition, understanding, and background world knowledge of the researcher. The initial search with the six sources resulted in 129 unique word pairs. But only 50 of them turned out to be relevant to our specific text, for an overall success rate of 39%. For the individual sources, the success rates were as follows: thesaurus 29%, WordNet 38%, semantic tagging 48%, association database 51%, LSA 56%, and MI 54%. This means that all three sources based on word associations (association database, LSA, and MI) had relatively false positives than the three sources based on word relationships (thesaurus, WordNet, and semantic tagging).

The full list of word pair matches found by the six sources, separated by relevance, along with the percentage of relevant pairs can be found in Table 1.

Table 1. Word Pairs found using the six automatable methods

	Thesaurus	**WordNet**	**Semantic Tagging**	**Small World of Words**	**LSA**	**COCA MI**
Word Pairs found that are <u>not</u> relevant to this specific text	air-looks art-culture audience-house capture-film chance-opening created-framed culture-music culture-polish culture-taste house-viewers leads-opens looks-sound looks-studied love-taste name-term	air-music art-film art-studied audience-chance audience-house canvas-name canvas-studied chance-opening created-mothers culture-polish culture-taste director-leads family-name house-studied leads-music looks-smell looks-sound looks-viewers name-term rain-sequence	art-camera art-culture camera-culture capture-offer created-framed created-opening eccentricity-usual framed-opening framed-productions opening-productions welcome-name	air-falling air-music art-created art-handmade art-music comforts-mothers comforts-love culture-music falling-love family-love framed-wood house-little house-mannered house-opens love-obsession love-mothers love-share music-sequence opens-welcome rain-roof reds-roof smell-taste	art-canvas canvas-scene chance-stay chance-welcome cozy-little director-share little-usual members-share mothers-nursery audience-opens camera-opens framed-opens framed-roof smell-taste usual-welcome	air-damp air-smell camera-capture damp-smell falling-love habitat-nursery mothers-nursery mothers-smoked rain-roof raindrops-roof raindrops-sound smell-taste viewers-welcome
Word Pairs found that <u>are</u> relevant to this specific text	audience-viewers damp-rain damp-rainy family-house music-sound scene-sequence	audience-viewers canvas-material created-film created-productions family-house film-productions film-scene film-sequence music-sound opening-opens opening-sequence rain-rainy	arcane-eccentricity canvas-wood cozy-handmade created-productions damp-raindrops director-film director-viewers family-mothers film-viewers reds-yellows	apple-reds audience-viewers autumnal-reds camera-film canvas-material comforts-cozy comforts-house created-productions damp-rain director-film egg-yolk-yellows family-members family-mothers film-productions film-scene film-sequence habitat-house handmade-wood house-roof material-wood music-sound polish-wood reds-yellows	audience-scene audience-writer camera-film created-writer damp-rain director-productions director-scene director-writer falling-rain family-members film-scene film-opens house-roof productions-scene falling-raindrops rain-raindrops damp-rainy rain-rainy reds-yellows	audience-viewers canvas-wood comforts-material damp-rain damp-rainy falling-rain falling-raindrops family-members film-productions handmade-wood material-wood opening-sequence rain-raindrops rain-rainy reds-yellows
Percentage of relevant pairs out of all pairs found	6 relevant / 21 total = .29	12 relevant /32 total = .38	10 relevant /21 total = .48	23 relevant /45 total = .51	19 relevant /34 total = .56	15 relevant /28 total = .54

The second phase of the analysis was to determine which of the six sources returned the most valid results for this text. This was done based on the assumption that an exceptionally well written text (such as the movie review used as data for this study) should have high connectedness and should be well-organized. In a well-organized text, the connected concepts should appear closer together than a random distribution throughout the text. To determine whether or not this is true for this specific text, the in-text distance between each of the words was computed, and a point biserial correlation was calculated between the word distance (1 to 72) and whether or not each of the possible 2,628 word-pair combinations was a match with that particular source. If the word-pair was a match, it was assigned a 1 and if it was not a match, it was assigned a 0. Therefore, a negative correlation would show that the matched word pairs are closer together in the sequence of the text than a random distribution. All six sources showed a negative correlation, meaning that the connected concepts appear relatively close together. The negative correlations are very small, however, ranging from the strongest correlation at -0.24 to the weakest at -0.02. Although these correlations are small, the number of possible word pairs is 2,628, and therefore, five of the six point biseral correlations are significant at $p<.05$ and three of them are significant at $p<.00001$. The correlations and p-values for all six sources can be found in Table 2.

In addition, the point biserial correlation results can be viewed as a validation triangulation (see Watson Todd, 2016) where the most valid source will have the largest negative correlation between the distance between the pair of words and whether or not the pair is a match from that source. The results show that the most valid source based on this test was the thesaurus, with a point biserial correlation of -0.14.

Table 2. Point Biserial Correlation between distances and relevance

Data Source	PBS Correlation
Thesaurus	-0.17**
WordNet	-0.04*
Semantic Tagging	-0.10**
Small World of Words	-0.24**
Latent Semantic Analysis	-0.02
COCA MI Scores	-0.12**

* $p < .05$ ** $p < .00001$

The third and final phase of this study was to compare the cumulative results from all six sources with the original manual analysis to determine whether or not the results from the automatable methods corroborated with the results produced from the researcher's intuition. To make this comparison, the number of times each word pair appeared in the six sources was calculated. Even though the maximum possible score is six sources, the highest actual score was achieved by three word pairs for only four sources, with the word pairs being: *audience-viewers*, *reds-yellows*, and *damp-rain*. Eight word pairs were found in three sources, 27 word pairs were found in two sources, 91 word pairs were found in only one source, and all the remaining 2,499 word pairs were not found in any sources. However, as discussed above, not all word pairs are relevant to our text. Looking at the word pair matches found in 1, 2, 3, or 4 sources, the number of sources correlated with whether or not the word pair match was relevant to our specific text. 100% of the pairs found in four sources agreed with the manual analysis, 88% of the pairs found in three sources agreed, 52% of the pairs found in two sources agreed, and only 29% of the pairs found in

only one source agreed. The point biserial correlation comparing number of sources for each of the unique 129 word pairs found in 1-4 sources and whether or not that word pair was relevant to our text was 0.39 which is significant at $p < 0.00001$. In other words, the more sources a word pair appeared in, the more likely the word pair was also identified as being a connected concept in our text. There was only one word pair (*bespoke-artisanal*) that we previously identified as being a word association in the text that was not found in any of the six sources. So, other than one word pair containing rare words, the six sources were able to match the connectedness of concepts in the text with the previous manual analysis based on our researcher intuition.

5. Conclusion

The purpose of this study was to investigate the use of six automatable methods for finding word relations and word associations in a sample text. Overall, the results showed that all six methods seem like reasonable ways to find relations and/or associations, as each method was able to find relevant word pairs (with the relevance being determined by the researcher as described above). The results above show that the methods that were primarily focused on associations outperformed those focused on relations. The association database, LSA, and MI returned more word pairs, as well as provided a higher percentage of word pairs which were relevant to our text than the thesaurus, WordNet, and semantic tagging.

In trying to determine which sources provided the best results, there are two issues worth considering: the number of word pairs found, and the percentage of word pairs which are relevant to our text. The difference in the total number of word pairs found could be attributed to several factors. The first is that thesauruses and WordNet are relations-only, while association databases, LSA, and MI were able to uncover not only associations, but also were able to find unique relations in our data such as *house-roof*, and even found connections that would be considered to be reiterations such as *rain-rainy*. Another factor which may influence the number of results is the way that the data in the different sources are created. Thesauruses and WordNet, on the one hand, may be limited because they are hand-made by human linguists and lexicographers who might not have the time or the resources to provide exhaustive lists of every possible connection that a word might have. Fully automated methods such as LSA and MI, on the other hand, can return thousands of word pair results, assuming that the underlying corpora are large enough. Therefore, sources such as LSA and MI will always be able to return a higher number of word pairs than a manually edited source such as a thesaurus.

For the issue of the percentage of word pairs which are relevant to our text, the relation sources gave a higher percentage of false positives than the association sources. This may be due to the fact that we did not take into account any weighting of different meanings of a word when using the thesaurus and WordNet. In other words, we treated relations created with obscure and rare meanings of a word to be just as important as those created with common meanings of the words. But both our text as well as the corpora underlying the LSA and MI methods are more likely to create connections using the common definitions of words. The examples earlier in this paper of *audience-house* and *audience-chance* are illustrative of this issue. The definition of *house* in the *audience-house* relation is the *audience* in a theatre, but our text uses the much more common definition of *house* being the building where a family lives. Likewise, for *audience-chance*, an *audience* is a *chance* to tell your side of the story in court, but our text uses the much more common definition of an *audience* being a group of people watching a performance. This issue of unweighted connections of polysemous words in a thesaurus and Wordnet might give

association-based sources an advantage in returning a higher percentage of relevant word pairs than purely relation-based sources.

From the perspective of which methods present a more valid representation of the connectedness in a text, four methods returned matches that had a highly significant negative point biserial correlation with the distance between concepts in the text, meaning that the methods are valid representations of the connectedness in the text. The Small World of Words association database was the most valid of the six sources, and the thesaurus was the second most valid. The two sources which showed the least amount of correlation were WordNet and LSA. Interestingly, these two methods also produced the highest ratios of word-pairs that were related to movies or movie making in general, such as *film-scene*. This brings to light a limitation of this part of the study. We only looked at word pairs and did not consider how these word pairs create word chains or word networks. Since this text is a movie review, it is not surprising that there are many movie-related concepts spread throughout the text. So even though film and scene are a distance apart of 42 concepts in our text (which we assumed to imply a poorly organized text) they are actually connected to many other movie terms over that long distance. So our assumption that only short distances imply well-organized text does not hold for words found in long chains or large networks in the text. Future research into the automation of word relation and word association identification should take this issue into consideration.

The results of this study also showed that the more often a word pair was identified in the six sources, the more likely that it was a valid connection in the context of our text. And since the only word pair that was not found was one containing rare words, then we can conclude that these methods in combination did an excellent job of identifying which words were relevant to our text. If we consider all the word pairs that were found in at least one source, it would reduce the number of potential word pairs from 2,628 to a much more manageable 129 pairs which then could be further analyzed manually by the researcher. These six methods therefore show promise for future full automation of the discovery of relations and associations in a text.

References

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

Corpus of Contemporary American English. (n.d.). Retrieved April 1, 2017 from http://corpus.byu.edu/coca/

Cruse, A. (2011). *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London:Longman.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Psychology Press.

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

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *Introduction to Latent Semantic Analysis. Discourse Processes*, 25, 259-284.

Latent Semantic Analysis @ CU Boulder. (n.d.) Retrieved April 1, 2017 from http://lsa.colorado.edu.

Lindberg, C.A. (2004). *The Oxford American Writer's Thesaurus*. USA: Oxford University Press.

Nelson, D.L., McEvoy, C.L. & Dennis, S. 2000. What is free association and what does it measure? *Memory and Cognition*. 28(6):887–899.

Princeton University. (2010). About WordNet. WordNet. Princeton University. Retrieved April 1, 2017 from http://wordnet.princeton.edu.

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

Small World of Words. (n.d.). Retrieved April 1, 2017 from https://smallworldofwords.org/en/project/home

UCREL. (n.d.). UCREL Semantic Analysis System (USAS). Retrieved April 1, 2017 from http://ucrel.lancs.ac.uk/usas/.

Watson Todd, R. (2013). Identifying new knowledge in texts through corpus analysis. *International Journal of Language Studies*. 7(4), 57-76.

Watson Todd, R. (2016). *Discourse Topics*. Amsterdam: John Benjamins.

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