Classification of Textual Genres using Discourse Information

Elnaz Davoodi, Leila Kosseim, Félix-Hervé Bachand, Majid Laali, and Emmanuel Argollo

Department of Computer Science & Software Engineering
Concordia University
Montreal, Canada
e_davoo@encs.concordia.ca, leila.kosseim@concordia.ca,
felixherve@gmail.com, m_laali@encs.concordia.ca,
emmanuel.argollo@gmail.com

Abstract. This papers aims to measure the influence of textual genre on the usage of discourse relations and discourse markers. Specifically, we wish to evaluate to what extend the use of certain discourse relations and discourse markers are correlated to textual genre and consequently can be used to predict textual genre. To do so, we have used the British National Corpus and compared a variety of discourse-level features on the task of genre classification.
The results show that individually, discourse relations and discourse markers do not outperform the standard bag-of-words approach even with an identical number of features. However, discourse features do provide a significant increase in performance when they are used to augment the bag-of-words approach. Using discourse relations and discourse markers allowed us to increase the F-measure of the bag-of-words approach from 0.796 to 0.878.

1 Introduction

Well-written texts are composed of textual units that are connected to each other via discourse relations. Such relations (e.g. CAUSE, CONDITION) communicate an inference intended by the writer and allow the creation of coherent connections between textual units. Discourse relations can be made explicit through discourse markers such as *but*, *since*, *because*, etc. or can be left implicit, when no explicit cue phrase is used to indicate the relation.

Previous work such as [26, 1, 18, 4] has shown a correlation between the use of discourse relations and certain textual dimensions, such as genre, level of formality and level of readability. For example, [26] has shown that the distribution of discourse relations in the PDTB corpus [19] is influenced by the textual genre; that is, texts from different genres tend to contain more of discourse relations than others.

The goal of this paper is to provide more insight on these preliminary investigations and measure the influence of textual genre on the usage of discourse relations and discourse markers on a larger scale. Specifically, we wish to evaluate to what extend the use of discourse relations and discourse markers are correlated to textual genre and consequently can be used to predict textual genre. To do so, we have used the British

National Corpus [2] which contains naturally occurring texts organized into various textual genres and compared a variety of discourse-level features on the task of genre classification.

2 Previous Work

In the literature, the term *genre* is often used to refer to slightly different concepts and is used in variety of domains such as linguistics, music, web documents, etc. [6]. In the context of written texts, [16] and [23] define textual genre using external criteria such as the type of intended audience, the communicative purpose, the activity type, etc. However, because these external criteria are difficult to detect automatically, efforts have been made to define textual genre using internal linguistic and structural properties which are easier to detect and measure [12, 10, 5].

Automatic genre classification is typically based on machine learning techniques that use structural and linguistic properties of texts. Previous work on automatic genre classification have generally followed two main approaches: linguistic analysis and frequency-based techniques. [11] used word and character statistics, part-of-speech (POS) frequencies as well as function word counts as features sets. They used the Brown corpus [7] and performed the classification in three genres. These features achieved a 73% accuracy on 4 genres. [12] investigated the influence of four sets of structural (e.g. POS frequencies), lexical (e.g. word frequencies), character level (e.g. punctuation and delimiter frequencies) and derivative features (i.e. ratio and variation of lexical and character-level features) on automatic genre classification. They pointed out that their feature set achieves a higher performance than [11]'s features; however there is no evidence that this improvement is statistically significant. In addition to focusing on a wide range of linguistic features, [8] used a bag-of-word document representation; however the data for this experiment was collected from the Internet which made the data set biased and evaluation was hard to reproduce. Instead of considering all of the words in the documents, [22] considered only the most common words in English as well as punctuation marks as a feature set and classified the Wall Street Journal articles into its four genres. The most common words in English were extracted from the BNC corpus. Four genres from the Wall Street Journal formed the corpus, but only 13 to 20 samples per genre were used to get the error rate of less than 7%.

To our knowledge, very little research has explored the influence of discourse-level properties on automatic genre classification. [26] has investigated the influence of textual genre on the usage of discourse relations within the PDTB corpus [19]. Although the corpus was rather small and skewed (1902 documents in the *news* genre, 104 documents in *essays*, 55 in *summaries*, and 49 in *letters*), she showed that the distribution of discourse relations is influenced by the textual genre. Moreover, it was pointed out that the genre appears to be a predictive feature for labelling discourse relations, especially when there is no lexical cue to signal the relation. More recently, [1] studied the usage of discourse relations across textual genres, and also across sub-topics of each genre. In their investigation, a wide range of corpora across various textual genres were used (e.g. [3], [24], [25], [19] and [20]). According to their corpus study, certain discourse relations are more likely to occur in certain textual genres, and further down, in certain

sub-topics of these specific genres.

As a follow-up to these recent works, we wanted to investigate if the differences in discourse-level usage noted by [26] and [1] were sufficient to be used as features for a textual genre classification task.

3 Methodology

3.1 The Corpus

In order to perform textual genre classification, we used a subset of the British National Corpus [2]. The British National Corpus (BNC) is a collection of English documents (100 million words) from various sources, both written and spoken. This corpus was selected because it is significant in size and is already divided into 8 textual genres. In this work, we only considered 4 of these genres, as we only deal with written documents (as opposed to spoken) and the definition of some classes is rather broad. The subset of the BNC that we used is composed of 2,179 documents (about 60,000,000 words) divided in 4 different textual genres:

1. Academic Prose (ACPROSE) is composed of documents containing specialized explanations in a specific field of study, such as research papers, academic theses, and studies. These types of documents are typically segmented into distinct sections, such as abstract, methodology, discussion, and results, each making use of various discourse structures.
2. Fiction (FICTION) contains documents that follow the general structure of a fictional story. It should be noted, that the structure of narrative fiction is not very strict.
3. News (NEWS) contains documents which retell series of events. These are typically news articles, recaps of sporting events, or political editorials.
4. Non-academic Prose and Biographies (NONAC) have a similar communicative purpose to academic prose. However, whereas academic prose targets audiences at the university-level, non-academic prose targets audiences with general knowledge. The NONAC genre is also mostly divided into sections, with the exception of biographies, with each section having its own discourse sructrure.

Table 1 summarizes the subset of the BNC used in our experiments. As Table 1 shows, the corpus is somewhat balanced both with respect to the number of documents (with NONAC being more frequent), and with respect to the number of words (with NEWS being generally shorter).

3.2 Discourse Features

In order to extract discourse-level information, the documents from the BNC subset were parsed using the PDTB End-to-End Discourse Parser [17]. Several publicly available discourse parsers could have been used (eg. [17, 14]). We chose the End-to-End

	Textual Genre				Total
	ACPROSE	FICTION	NEWS	NONAC	
Number of documents	497	452	486	744	2,179
Number of words	15,715,469	15,806,443	4,300,672	24,064,370	59,886,954
Average nb of words / document	31,621	34,970	19,158	32,345	27,484

Table 1. Statistics on the BNC sub-Corpus Used

parser as it is the most commonly used parser and provides local discourse-level information such as the type of discourse relations (i.e. *implicit* or *explicit*), the name of the relation (known as its *sense*) and the discourse marker when applicable. The End-to-End parser uses the PDTB [19] set of discourse relations organised into 3 levels of granularity: 4 relations at level 1, sub-divided into 12 relations at level 2 themselves sub-divided into 23 relations at level 3. For the purpose of this work, we considered the 12 relations[1] at level 2 for which the End-to-End parser achieves the best performance. In addition, to tag discourse markers, the End-to-End parser uses the set 100 discourse markers from the PDTB.

Several features were used in order to evaluate the influence of discourse-level information on textual genre:

Discourse Relations (DR) at level-2 of the PDTB were used as the first set of features. As indicated above, these were extracted from the documents using the End-to-End Discourse parser [17]. For this feature, we used all the relations identified by the End-to-End parser regardless of how these were realized in the documents: *explicitly* through a discourse marker or *implicitly*[2]. This gave rise to 12 features.
For the value of each DR feature, we used the Log Likelihood ratio as defined by [21]. This measure was used as it indicates how many times each DR is more likely to occur in a specific genre as opposed to another.

Discourse Markers (DM) can be used to signal several discourse relations. For example, the marker *since* can signal a TEMPORAL or a CAUSAL relation. According to [15], markers can signal on average 3.05 different relations in the PDTB. The DR feature above takes into account both explicitly stated relations (those signalled via a discourse marker), as well as implicit relations (those that are not signaled by a marker). In order to focus only on explicit relations and to minimize the effect of mislabelled discourse relations, we also used discourse markers (as tagged by the End-to-End parser) as a feature set. Previous work have used different sets of discourse markers (e.g. [13]); however, to ensure consistency with the previous feature, we used the list of 100 discourse markers used in the PDTB corpus [19].
Here also, we used the Log Likelihood ratio to calculate how many times more likely each DM is in each genre. Each document is represented as a vector of 100

[1] which include ASYNCHRONOUS, SYNCHRONOUS, CAUSE, CONDITION, CONTRAST, CONCESSION, CONJUNCTION, INSTANTIATION, RESTATEMENT, ALTERNATIVE, EXCEPTION and LIST.

[2] We experimented with using only explicit relations and only implicit relations, but the results were not conclusive.

features (one for each DM), and the value of each feature is its Log Likelihood ratio.

It is important to note that the discourse markers that we used, actually mark a relation. Indeed, some cue phrases (such as *and*) may be used to signal a discourse relation (e.g. CONJUNCTION) or may be used in a non-discourse marking role. Section 4.3 analyses the difference between the use of discourse markers (that do signal a discourse relation) as opposed to cue phrases (that may not signal a relation).

Bag-of-words (BOW) To compare the effectiveness of the above features, we used a standard bag-of-words approach as a baseline. Words were extracted after case-folding, stemming, and digit and punctuation removal. In addition, for comparative purposes, we also used the Log Likelihood ratio of words across the four genres and only considered the words that had a Log Likelihood ratio up to 100 times less than the highest Log Likelihood ratio. This gave rise to 2,233 features.

Table 2 summarizes statistics on the discourse-level features extracted from our BNC sub-corpus. As Table 2 shows, the NEWS genre seems to contain significantly less DRs, but recall from Table 1 that it also contains less words. On average, this genre contains more DRs (1 DR every 7.30 words) compared to the other genres (1 DR every 13 or 18 words for the other genres). Another interesting remark is that the NEWS genre seems to use more discourse markers (on average, 1 word out of 22 is a marker) whereas the other genres have a marker every 44 to 50 words.

	Textual Genre				**Total**
	ACPROSE	FICTION	NEWS	NONAC	
Number of Discourse Relations (DR)	836,861	1,236,677	591,463	1,335,213	4,000,214
Number of Discourse Markers (DM)	343,170	352,772	197,395	498,281	1,391,618
Number of Cue Phrases (CP)	1,190,664	1,148,681	285,221	1,804,345	4,428,911
Ratio nb of words / DR	18.78	12.82	7.30	18.18	15.15
Ratio nb of words / DM	47.61	45.45	22.22	50.00	43.48

Table 2. Discourse-level Features in the BNC sub-Corpus Used

4 Results and Analysis

To perform the classification task, we used 3 classifiers provided by WEKA [9]: Multinominal Naïve Bayes, Decision Tree, and Random Forest. The first two classifiers were used as a baseline; while the Random Forest was investigated for its properties of reducing overfitting.

4.1 Initial Results

Table 3 shows the results obtained in terms of precision, recall, and F-measure using 10-fold cross validation. The last column of Table 3 indicates if there is a statistically

significant decrease (⇓) or no difference (=) in F-measure compared to the bag-of-words (BOW) model.

As Table 3 shows, the best results are consistently obtained using the bag-of-words features, regardless of the classifier used. This set of features is however, much larger than the others (2,233 vs 100 and 12). The performance of discourse markers (DM), with only 100 features, is very close to that of the BOW. In the case of the Random Forest classifier, it even achieves the same performance as BOW[3]. Finally, discourse markers (DM) achieve a better F-measure with all three classifiers, than discourse relations (DR).

Classifier	Features	# Features	P	R	F	Stat. Sign.
Naïve Bayes	BOWtop2233	2233	0.761	0.745	0.733	
	DM	100	0.682	0.662	0.653	⇓
	DR	12	0.550	0.517	0.511	⇓
Decision Tree	BOWtop2233	2233	0.757	0.746	0.748	
	DM	100	0.695	0.695	0.695	⇓
	DR	12	0.629	0.629	0.629	⇓
Random Forest	BOWtop2233	2233	0.816	0.798	0.796	
	DM	100	0.797	0.800	0.797	=
	DR	12	0.717	0.717	0.715	⇓

Table 3. Initial Results of the Classification Task

4.2 Influence of the Feature Size

As shown in Table 3, the BOW features achieve the best results; however, it has two major drawbacks: First, it constitutes a much larger feature set than the other two approaches, and second, the actual words used as features need to be identified for each corpus and hence these features are tailored for the corpus at hand. On the other hand, discourse relations and discourse markers both constitute a small feature set and the features are fixed for all corpora. To investigate this further, we performed additional experiments, but this time, we reduced the number of features so as to be at par across all experiments. Specifically, we took:

1. BOWtop12: the top 12 most discriminating words.
2. BOWrandom12: 12 random words, to be used as a baseline.
3. BOWtop100: the top 100 most discriminating words.
4. BOWrandom100: 100 random words, to be used as a baseline.
5. DMtop12: the top 12 most discriminating discourse markers.

The results are shown in Tables 4 and 5. Not surprisingly, random words always achieve the lowest performance (equivalent to picking the most frequent genre). Most importantly, the tables show that even when the cardinality of the features sets are identical, the BOW approach still outperforms discourse relations and discourse markers.

[3] The difference between the two F-measures is not statistically significant.

Classifier	Features	# Features	P	R	F	Stat. Sign.
Naïve Bayes	BOWtop12	12	0.679	0.617	0.610	
	DR	12	0.550	0.517	0.511	⇓
	DMtop12	12	0.598	0.607	0.573	⇓
	BOWrandom12	12	0.347	0.250	0.227	⇓
Decision Tree	BOWtop12	12	0.689	0.682	0.682	
	DR	12	0.629	0.629	0.629	⇓
	DMtop12	12	0.622	0.622	0.619	⇓
	BOWrandom12	12	0.322	0.349	0.266	⇓
Random Forest	BOWtop12	12	0.729	0.720	0.717	
	DR	12	0.717	0.717	0.715	⇓
	DMtop12	12	0.675	0.669	0.661	⇓
	BOWrandom12	12	0.457	0.351	0.266	⇓

Table 4. Results of Classification Tasks using 12 Features

Classifier	Features	# Features	P	R	F	Stat. Sign.
Naïve Bayes	BOWtop100	100	0.723	0.705	0.681	
	DMtop100	100	0.682	0.662	0.653	⇓
	BOWrandom100	100	0.425	0.411	0.382	⇓
Decision Tree	BOWtop100	100	0.746	0.737	0.739	
	DMtop100	100	0.695	0.695	0.695	⇓
	BOWrandom100	100	0.532	0.511	0.504	⇓
Random Forest	BOWtop100	100	0.818	0.805	0.805	
	DMtop100	100	0.797	0.800	0.797	⇓
	BOWrandom100	100	0.544	0.525	0.519	⇓

Table 5. Results of Classification Tasks using 100 Features

4.3 Discourse Markers versus Cue Phrases

Recall from Section 3.2, that to be considered a discourse marker, cue phrases need to mark a discourse relation. For example, the cue phrase *since* was considered as a discourse marker only if it marked a discourse relation (i.e. CAUSE). If the *since* did not mark a relation, as in:

(1) Equitable of Iowa Cos., Des Moines, had been seeking a buyer for the 36-store Younkers chain *since* June, when it announced its intention to free up capital to expand its insurance business [4].

then it was not counted as a feature. The intuition behind this was to avoid adding noisy features as these cue phrases are often grammatical words. The disadvantage, however, is that a discourse parser is required to parse the documents in advance to identify discourse relations and markers. To evaluate if the use of such a discourse parser was really necessary, we compared the use of both cue phrase (CP) and discourse markers (DM). Hence we performed the same genre classification task again but replaced discourse markers (DM) with their corresponding cue phrases (CP) without verifying if

[4] The example is taken from the PDTB [19].

these were used to mark a discourse relation of not. As shown in Table 6, using all cue phrases does not provide as much information as using only the cue phrases that signal a discourse relation. This seems to show that it is not the cue phrase per se that is a discriminating feature, but its discourse usage.

Classifier	Features Used	Nb. Features	P	R	F	Stat. Sign.
Naïve Bayes	DM	100	0.682	0.662	0.653	
	CP	100	0.544	0.493	0.493	⇓
Decision Tree	DM	100	0.695	0.695	0.695	
	CP	100	0.560	0.559	0.549	⇓
Random Forest	DM	100	0.797	0.800	0.797	
	CP	100	0.606	0.603	0.595	⇓

Table 6. Results of Classification using Discourse Markers (DM) vs Cue Phrases (CP).

4.4 Feature Combination

As Sections 4.1 and 4.2 showed, regardless of the number of features, discourse features alone do not achieve the performance of BOW. However, since these features are complementary, we tried to combine them to see if discourse features could somehow improve the BOW model. We used the best performing BOW approach (BOWtop2233) and augmented it with discourse relations only, discourse markers only and both discourse relations and markers.

As Table 7 shows, augmenting the BOW model with all discourse features (DM + DR) increases the F-measure of all classifiers significantly. For example, the F-measure of the Random Forest classifier increases from 0.796 to 0.878 with discourse features. Note that this increase in performance is significant and only requires the addition of 112 features (12 DR + 100 DM). The difference between the effect of discourse markers and discourse relations is not significant - however, because the two features measure essentially the same linguistic phenomena, the use of both features may not be necessary. The best performance was in fact achieved using a Random Forest classifier with the BOW model and only the addition of discourse markers (F=0.884). Considering that discourse relations are made up of only 12 features, they might constitute a good choice to augment the standard BOW approach.

5 Conclusion and Future Work

This paper aimed to measure the influence of textual genre on the usage of discourse relations and discourse markers. Specifically, we evaluated to what extend the use of discourse relations and discourse markers are correlated to textual genre and consequently can be used to predict textual genre. To do so, we have used a subset of the British National Corpus and compared a variety of discourse-level features on the task of genre classification.

Classifier	Features	# Features	P	R	F	Stat. Sign.
Naïve Bayes	BOWtop2233	2233	0.761	0.745	0.733	
	BOWtop2233+DR	2245	0.809	0.807	0.803	⇑
	BOWtop2233+DM	2333	0.806	0.804	0.801	⇑
	BOWtop2233+DM+DR	2345	0.806	0.804	0.801	⇑
Decision Tree	BOWtop2233	2233	0.757	0.746	0.748	
	BOWtop2233+DR	2245	0.809	0.810	0.809	⇑
	BOWtop2233+DM	2333	0.811	0.811	0.811	⇑
	BOWtop2233+DM+DR	2345	0.810	0.810	0.810	⇑
Random Forest	BOWtop2233	2233	0.816	0.798	**0.796**	
	BOWtop2233+DR	2245	0.879	0.879	0.870	⇑
	BOWtop2233+DM	2333	0.884	0.884	0.884	⇑
	BOWtop2233+DM+DR	2345	0.878	0.878	**0.878**	⇑

Table 7. Final Results of the Classification Task

The results show that individually, discourse relations and discourse markers do not outperform the standard bag-of-words approach even when the number of features is identical. However, discourse features do provide a significant increase in performance when they are used to augment the bag-of-words approach. Using discourse relations and discourse markers allowed us to increase the F-measure of the BOW approach from 0.796 to 0.878. This seems to show that discourse information models a linguistic phenomenon which the bag-of-words does not. The bag-of-words approach can model well textual topic, however for textual genre, this approach is not enough, and can be complemented by discourse information.

Our investigations also showed that although the BOW approach achieves a better performance, the actual words used as features need to be identified for each corpus and hence these features are tailored for the corpus at hand. On the other hand, discourse relations and discourse markers both constitute a small feature set (12 relations and 100 markers) and the features are fixed for all corpora. Using only these features, the Naïve Bayes and Decision Tree approaches achieve lower results than the the BOW approach; however, with the Random Forest classifier, discourse markers alone achieve results that are statistically equivalent to the tailored BOW approach. If tailoring the feature set to the corpus is not an option, discourse markers constitute a very good alternative for genre classification.

Another interesting result is the fact that not all cue phrases are discriminating for genre classification. Our results show that using cue phrases (*since*, *and*, *because* ...) that actually mark a discourse relation produce a higher performance than using all cue phrases regardless of whether they signal a discourse relation or not. This is not surprising, as most cue phrases are grammatical words that have a low discriminating power. The fact that discourse markers are good indicators of textual genre may indicate that their usage to mark discourse relations is different across genres.

Our work has focused on the analysis of 4 genres of the British National Corpus. An obvious question to investigate is the validity of our results on other corpora. The

Brown corpus [7], for example, also provides samples of different genres and could be used to validate our results.

Another very interesting question is to investigate how the conclusions drawn from our work can be used to improve discourse parsing. Our experiments seem to show that different genres have different discourse properties that are significant enough to be used as a basis for textual genre classification. Therefore, as [26] noted, it might be the case that knowing the textual genre of a document could be an interesting feature to improve discourse parsing. Knowing, for example, that the text being parsed is an academic prose as opposed to a work of fiction might be useful to properly label a text span with a discourse relation as opposed to another.

Acknowledgement

The authors would like to thank the anonymous reviewers for their feedback on the paper. This work was financially supported by NSERC.

References

[1] Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. An investigation on the influence of genres and textual organisation on the use of discourse relations. In *Proceeding of the 15th International Conference of Computational Linguistics and Intelligent Text Processing (CICLing), LNCS-volume 8404*, pages 454–468. Springer, 2014.

[2] BNC Consortium. The British National Corpus, version 3 (BNC XML Edition). 2007. http://www.natcorp.ox.ac.uk/.

[3] Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. *RST Discourse Treebank*. Linguistic Data Consortium, LDC2002T07, University of Pennsylvania, 2002.

[4] Elnaz Davoodi and Leila Kosseim. On the influence of text complexity on discourse-level choices. *International Journal of Computational Linguistics and Applications*, 6(1):27—-42, 2015.

[5] Chengyu Alex Fang and Jing Cao. *Text Genres and Registers: The Computation of Linguistic Features*. Springer, 2015.

[6] Aidan Finn and Nicholas Kushmerick. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11):1506–1518, 2006.

[7] Winthrop Nelson Francis. *A manual of information to accompany A standard sample of present-day edited American English, for use with digital computers*. Department of Linguistics, Brown University, 1971.

[8] Luanne Freund, Charles LA Clarke, and Elaine G Toms. Towards genre classification for IR in the workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, pages 30–36, New York, NY, USA, 2006.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD explorations*, 11(1):10–18, 2009.

[10] Jussi Karlgren. Stylistic experiments in information retrieval. In *Natural Language Information Retrieval*, volume 7 of *the series Text, Speech and Language Technology*, pages 147–166. Springer, 1999.

[11] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics (ACL) - Volume 2*, pages 1071–1075, Las Cruces, USA, 1994.

[12] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL)*, pages 32–38, Madrid, Spain, 1997.

[13] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, 1996.

[14] Majid Laali, Elnaz Davoodi, and Leila Kosseim. The CLaC Discourse Parser at CoNLL-2015. In *CoNLL 2015*, pages 56–60, Beijing, China, 2015.

[15] Majid Laali and Leila Kosseim. Inducing Discourse Connectives from Parallel Texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 610—619, Dublin, Ireland, 2014.

[16] David Yw Lee. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Technology*, 5:37–72, 2001.

[17] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 1:1–34, 2012.

[18] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195, Honolulu, October 2008.

[19] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Morocco, 2008.

[20] Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The biomedical discourse relation bank. *BMC Bioinformatics*, 12(188), 2011.

[21] Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6, Hong Kong, 2000.

[22] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th Conference on Computational Linguistics (ACL) - Volume 2*, pages 808–814, 2000.

[23] John Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

[24] Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genova, Italy, 2006.

[25] Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 158 – 161, Stanford University, CA, 2004.

[26] Bonnie Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-AFNLP: Volume 2)*, pages 674–682, Suntec, Singapore, August 2009.