On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse

Judith Eckle-Kohler[1] **Roland Kluge**[3] **Iryna Gurevych**[1,2]
[1] UKP Lab, Technische Universität Darmstadt
[2] UKP Lab, German Institute for Educational Research
[3] Real-Time Systems Lab, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

This paper presents a study on the role of discourse markers in argumentative discourse. We annotated a German corpus with arguments according to the common claim-premise model of argumentation and performed various statistical analyses regarding the discriminative nature of discourse markers for claims and premises. Our experiments show that particular semantic groups of discourse markers are indicative of either claims or premises and constitute highly predictive features for discriminating between them.

1 Introduction

The growing field of argumentation mining in NLP develops methods to automatically analyze argumentative discourse for enhanced Information Extraction.

Terminology: We understand argumentation as a rhetorical arrangement of *arguments* with the intent to convince or persuade the reader of a particular state of affairs. Following previous work in argumentation mining (e.g., (Palau and Moens, 2009; Florou et al., 2013; Peldszus and Stede, 2013a; Stab and Gurevych, 2014)), we define an *argument* as the pairing of a single *claim* (an arguable text unit) and a (possibly empty) set of *premises*, which each either support or attack the claim (Besnard and Hunter, 2008). We subsume claims and premises under the term *argument unit* (AU).

Discourse markers in argumentative discourse: Since an argumentation line can only be captured in the context of a coherent text, argumentation mining is closely related to automated discourse analysis (Cabrio et al., 2013), which aims at identifying discourse functions of text segments, and discourse relations (DRs) between adjacent text segments (Webber et al., 2012). Often, so-called discourse markers (DMs) are used to signal DRs. The following example shows that DMs act as lexical markers in argumentative discourse as well: the DM *however* (marking the DR *contrast*) positions the claim in (1) in the overall argumentation line, while in (2), the DMs *as* (marking *reason*) and *also* (marking *elaboration*) connect the premises with each other and with the claim.

> (1) *However*, staying down is pointless from a pedagogic point of view.
>
> (2) *As* the students get frustrated, their performance generally does not improve. *Also*, the point of repeating all courses because of only one or two bad grades is arguable.

DMs belong to the word classes of conjunctions and adverbs (also called discourse particles) and are semantically characterized in traditional grammar books. The correspondence between DM semantics and DR semantics has received considerable attention in previous research in linguistics, most of which is based on corpora annotated with DRs (Carlson et al., 2003; Wolf and Gibson, 2005; Prasad et al., 2008). In contrast, the role of DMs as potential lexical signals in *argumentative* discourse is not well-understood, yet. While Stab and Gurevych (2014) used DMs as features for classifying AUs into different types, they did not analyze the semantics of DMs with respect to AU classification or considered different DM resources.

As far as we are aware, there is no prior work performing a detailed investigation on the role of DMs as lexical signals for discriminating between the two fundamental argumentative roles of AUs, i.e., between claims and premises.

Our contribution: In this paper, we address this gap by analyzing the role of DMs for the automatic discrimination of claims and premises in a

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2236–2242,
Lisbon, Portugal, 17-21 September 2015.

new dataset of argumentative texts annotated with arguments. More specifically, we (i) present the results of an annotation study that we performed to annotate a German dataset of news documents with arguments, and (ii) performed a detailed investigation of the role of DMs in our annotated dataset which highlights correspondences between DM semantics and semantic properties of claims and premises, and shows that DMs are highly predictive features for automatically discriminating premises and claims.

2 Related Work

Related to our work are prior investigations (i) on the relationship of DMs and DRs, and (ii) on the role of DMs in classification tasks.

2.1 Relation of DMs and DRs

Previous work on the relation between DMs and DRs is mostly based on corpora annotated with DRs (Wolf et al., 2004; Taboada, 2006; Prasad et al., 2008), most notably the Penn Discourse Treebank (PDTB) for English (Prasad et al., 2008). The PDTB is annotated with DRs, such as *contrast* or *result*, and the corresponding DMs, even if they were not realized in the text. For instance, the *contrast* relation can be expressed by the DMs *however* and *but*. DRs that are lexically signaled by DMs in the text are called *explicit* DRs. Some DMs are highly polysemous, e.g., *while* appears in 12 DRs in the PDTB.

Asr and Demberg (2013) analyzed the DMs and their corresponding DRs annotated in the PDTB and addressed the question, which information is conveyed by discourse connectives in the context of human sentence processing, i.e., how they contribute in the process of inferring a particular DR.

Taboada (2006) performed a corpus-based analysis of DRs annotated in the Rhetorical Structure Theory (RST) Discourse Treebank (Carlson et al., 2003). The most frequent relation in the RST Discourse Treebank is *concession*, and this relation also received particular attention in the corpus linguistics literature: Taboada and Gómez-González (2012) present a corpus-based comparative study of DMs that express *concession* across English and Spanish in different genres. A classification of DMs signaling *concession* across English and German is presented by Grote et al. (1997). They also point out the importance of *concession* in argumentative discourse: DMs expressing *concession* are often used to introduce counter-arguments in an argumentation line.

2.2 DMs in Classification

There is previous work in sentiment classification and argumentation mining using DMs as features, as well as work in predicting DMs for natural language generation tasks, such as abstractive summarization. Taboada et al. (2011) successfully employed discourse particles as features for the calculation of polarity scores in automated sentiment analysis. They focused on particles acting as intensifiers, which modify the semantic intensity of the lexical item they refer to.[1] Mukherjee and Bhattacharyya (2012) demonstrated that using discourse connectives as features in a system for sentiment classification of Twitter posts significantly improves classification accuracy over state-of-the art systems not considering DMs as features.

In argumentation mining, Stab and Gurevych (2014) experimented with different types of features, including DMs from the PDTB annotation guidelines, to classify text units into the classes non-argumentative, major claim, claim, and premise. While the PDTB DMs appeared to be not helpful for discriminating between argumentative and non-argumentative text units, they were useful to distinguish between the classes premise and claim.

Patterson and Kehler (2013) describe a classification model, trained and evaluated on PDTB data, for predicting whether or not a DR is signaled by an explicit DM. The most predictive features in their model are discourse level features that encode dependencies between neighboring DRs given by the overall discourse structure. Other highly predictive features turned out to be the DMs themselves, because DMs vary as to their rates of being realized explicitly.[2]

3 Annotation Study

For our study, we used a dataset of 88 documents in German – mainly news (ca. 83% of the documents) – from seven current topics related to the German educational system (e.g., mainstreaming, staying down at school). The documents were manually selected from a focused crawl and the top 100 search engine hits per topic (Vovk, 2013).

[1] Amplifiers such as *very* increase the semantic intensity, while downtoners such as *slightly* decrease it.

[2] In the PDTB, DMs are also given for implicit DRs.

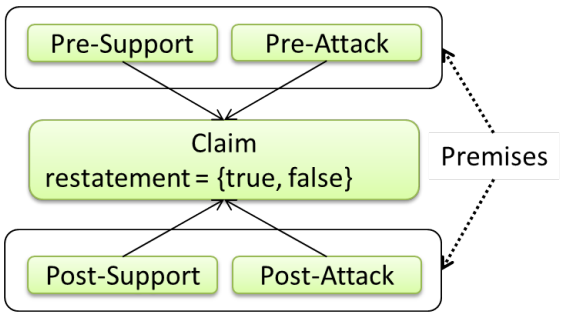

Figure 1: Claim-Premise scheme.

3.1 Annotation Scheme

Since argumentation theory offers a wide range of models, we performed a pre-study on a held-out development set to develop the annotation scheme. We found that most arguments consisted of adjacent claims and premises, related by a support or attack relation; premises and claims were rarely nested. Therefore, we decided to use a simplified claim-premises scheme (see Figure 1), using the terminology set up in Section 1.

Our scheme models an argument as a linked set of AUs and encodes the argumentative support and attack relations by assigning an argumentative role to *non-nested* (e.g., adjacent) AUs. The non-nested structure in combination with the *pre-* and *post-* prefixes of premises (indicating if a premise precedes or follows the claim) makes sure that all premises are correctly attached to their respective claims.

For claims, we also annotated whether the claim is a restatement. In total, we distinguish six argumentative roles as shown in Figure 1: *claim, restatement, pre-claim support, post-claim support, pre-claim attack and post-claim attack*. Annotators could freely select annotation boundaries, but we encouraged them to annotate clauses or sentences.

In the main study, three annotators annotated the remaining 80 documents (3 863 sentences) belonging to six topics. On average, each annotator marked 1 708 AUs (53 % premises, 47 % claims) and 783 arguments (2.2 AUs per argument). An average claim spans 1.1 sentences, whereas an average premise spans 2.2 sentences. On average, 74 % of the tokens are covered by an AU, indicating that the documents are highly argumentative.

	$A_{o,t}$	κ_t	$A_{o,s}$	κ_s	α_u
all	61.0	44.2	60.9	45.2	40.2
prem. / claims	62.7	45.1	62.6	46.3	41.7
AU / non-AU	79.0	50.0	77.0	50.0	56.6

Table 1: Inter-annotator agreement scores (percentages): A_o – observed agreement, κ_t– token-based kappa; κ_s– sentence-based kappa; α_u– unitized alpha.

3.2 Inter-annotator Agreement

Like in other discourse annotation tasks, there is no straightforward way to compute inter-annotator agreement (IAA) due to free annotation boundaries, see for instance (Miltsakaki et al., 2004; Wilson and Wiebe, 2003). We selected sentence-based *kappa* κ_s, token-based *kappa* κ_t and *Krippendorff's unitized alpha* α_u as IAA metrics.[3] The *token-based kappa* κ_t metric treats each token as annotation item and, thereon, calculates Fleiss' kappa (Fleiss, 1971). The tokens are labeled with the IOB scheme, i.e., every token is annotated as inside an AU (I), beginning of an AU (B) or outside an AU (O). In contrast, *Krippendorff's unitized alpha* α_u operates on whole annotation spans instead of isolated tokens. For comparison, we also calculated the observed agreement A_o (token-based and sentence-based).

Table 1 summarizes the IAA scores for three scenarios: (i) considering all six labels and non-AUs, (ii) only premise, claim and non-AUs, and (iii) AUs versus non-AUs. The IAA scores are in line with previous results: Peldszus and Stede (2013b) report κ=38.4 % and α=42.5 % for their sentence-level annotation study, which used artificially created documents.

4 Experiments

This section describes the experiments we performed to understand the role of DMs for the automatic discrimination of claims and premises. Since the IAA was not substantial, we performed all experiments on three separate datasets, one per annotator.

[3] We calculated the metrics using DKPro Statistics (Meyer et al., 2014).

4.1 Semantically Categorized DMs

Our first set of experiments aimed at understanding the correspondence of DM semantics and semantic aspects of claims and premises. For this, we used semantically characterized lists of DMs compiled from three different sources: (i) 28 semantically categorized particles from a German grammar (Helbig and Buscha, 1996, pp. 481–484), (ii) 51 discourse connectives based on a manual translation of the DMs in Appendix B of the PDTB annotation guidelines, and (iii) a large German lexicon of about 170 DMs called DiMLex[4] (Stede, 2002).

First, we applied a two-sample statistical test to find out if the two classes *claim* and *premise* are significantly different regarding the number of occurrences of individual DMs and semantic groups of DMs (based on the semantic information given in our lists). We chose Fisher's exact test (Fisher, 1932), a non-parametric randomization test that makes no assumptions about the underlying probability distribution of the DMs.[5] Lacking DR annotations, we counted surface word forms of single word and continuous multi-word DMs in the AUs annotated in our dataset.[6] For each semantically categorized DM, we computed a contingency table containing the number of observed occurrences in the two samples of claims and premises. For each semantic group of DMs given in DiMLex and PDTB, we calculated the per-group contingency table by adding up the contingency tables for each DM in that particular group.

The results of the significance tests revealed both single DMs and semantic groups of DMs that occur with significantly different frequency in claims or premises. The following individual DMs and semantic groups of DMs appeared to be indicative for claims and premises (note that the sentence-initial variants are counted separately):

Claims: four DMs expressing *result, comparison, contrast*:

- *also* (therefore), *Doch* (However), *jedoch* (though), *sondern* (but), as well as the amplifier *ganz* (quite);
- semantic groups of DMs from PDTB: *comparison* (expressing *concession* and *contrast*) and *result*.

Premises: three DMs expressing *cause, reason, elaboration, alternative*:

- *Denn* (As), *oder* (or), *und* (and), as well as the downtoner *etwa* (roughly);
- semantic groups of DMs from PDTB: *alternative* (e.g., *or*), *reason* (e.g., *because*), and from DiMLex, the group *sequence* (e.g., *then*).

In the group of high-frequent but non-indicative DMs are DMs expressing *elaboration* (*Auch* (Also), *Und* (And), *So* (Therefore)), *sequence* (*dann* (then)), and *contrast* (*aber/Aber* (but/But)), as well as some highly ambiguous particles (e.g., *immer* (always), *schon* (already)).

Second, we ranked the DMs according to their Information Gain (IG) using Weka (Hall et al., 2009). For this, we mapped each DM to a boolean feature indicating if it is present in an AU or not. We considered all the DMs from the three resources as features and ranked them by IG separately for each annotator. The resulting ranking revealed additional DMs indicative for premises, e.g. further DMs expressing *elaboration* and the downtoner *nur* (only).

Our findings are in accordance to the ability of a claim to act as conclusion or result of an argument, and to the role of premises as providing support for a claim by giving reasons and elaborating on them. Moreover, we found that particular intensifying discourse particles play an important role in discriminating claims and premises: Downtoners seem to be significant for premises, amplifiers for claims. Finally, semantic groups of DMs expressing *concession* appeared to be significant for claims, since claims often introduce a counter-argument in the overall argumentation structure (Grote et al., 1997).

4.2 Classification

The significance test and the IG ranking evaluated the DMs in isolation. In order to examine the predictiveness of all DMs *in combination* for claims vs. premises, we used a classification model. Given an AU, the model predicts its argumentative role, i.e., claim vs. premise. In this experiment, we used a list of single word DMs, extracted from the three DM resources described above, as boolean features (DMres).

[4] https://github.com/discourse-lab/dimlex

[5] We used the implementation given in http://stat.ethz.ch/R-manual/R-patched/library/stats/html/fisher.test.html and a p-value of 0.05.

[6] Sentence initial DMs were counted separately to capture DRs being signaled by a sentence initial position.

For comparison, we used two sets of data-driven DMs extracted from the Tiger corpus (Brants et al., 2004), a German newspaper corpus: one set containing the 350 most frequent conjunctions and adverbs (DMtiger), and another set containing the 300 most frequent non-open-class words (NOCtiger, excluding the word classes of nouns, main verbs and adjectives).[7] In addition, we considered the top 1 800 unigrams (with minimum frequency 5) as baseline features.

We trained three Machine Learning algorithms (Naive Bayes (NB), Random Forests (RF) and Support Vector Machine (SVM)) on the three datasets annotated by the annotators A1, A2 and A3, using 10-fold cross-validation and the DKPro TC framework (Daxenberger et al., 2014). For the classification experiments, we used all 88 documents in the annotated corpus (including the documents from the pre-study).

Table 2 summarizes the results. All classifiers show significant improvement compared to the majority class baseline (MC), indicating that DMs might be useful as predictive features for discriminating claims and premises. The DMs extracted from Tiger did not improve the performance consistently for all three datasets, showing that the coverage of the manually compiled DMs is good. Using the NOCtiger set, however, improved the performance by up to 4 pp., compared to DMres; the improvement of NOCtiger over DMres is statistically significant for NB across all datasets A1 - A3 (using Fisher's exact test and a p-value of 0.05). These results suggest that not only DMs, but also other function words, as well as auxiliaries and modals, play an important role in discriminating claims and premises.

While the unigram baseline also outperforms DMres, it is on par with NOCtiger (no significant improvement for any of NB, RF, SVM across the three datasets), but at the cost of a much larger feature space and a model less able to generalize to other datasets from the news domain.

	A1	**A2**	**A3**
MC	53.04	52.05	51.71
DMres-NB	64.74	65.21	64.53
DMres-RF	64.89	63.61	63.50
DMres-SVM	68.50	66.31	67.06
DMtiger-NB	67.81	65.65	66.65
DMtiger-RF	64.65	61.12	63.44
DMtiger-SVM	70.37	65.59	67.57
NOCtiger-NB	70.86	68.80	68.87
NOCtiger-RF	67.06	64.71	65.51
NOCtiger-SVM	71.80	68.25	68.41
unigram-NB	72.79	69.69	69.60
unigram-RF	70.32	68.75	68.15
unigram-SVM	71.21	69.97	71.14

Table 2: Accuracy (percent) using different feature sets.

5 Summary

Our goal was to understand some of the lexical-semantic characteristics of claims and premises with a focus on DMs acting as lexical signals for the argumentative role of AUs. Such insights into the way claims and premises are signaled by lexical markers can be exploited not only for the automatic analysis of argumentative discourse, but also for language generation tasks such as creating abstractive summaries of argumentative documents.

We investigated the role of a large set of DMs in argumentative discourse based on a German dataset annotated with arguments and identified semantic groups of DMs that are indicative of either claims or premises. These semantic groups also shed light on semantic aspects of claims and premises. Our classification model shows that DMs are important features for the discrimination of claims and premises. In order to foster further research on DMs in argumentative discourse, we publicly released the annotation guidelines, as well as the semantically categorized DM lists used in our experiments.[8]

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, by the German Institute for Educational Research (DIPF), and by the German Research Foundation under grant No. GU 798/17-1. We thank the anonymous reviewers for their valuable comments. We also thank Georg Weber for his contributions to the annotation study.

[7] We used DKPro Core (Eckart de Castilho and Gurevych, 2014) to pre-process Tiger.

[8] https://www.ukp.tu-darmstadt.de/data/argumentation-mining/argument-annotated-news-articles/

References

Fatemeh Torabi Asr and Vera Demberg. 2013. On the Information Conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In João Leite, TranCao Son, Paolo Torroni, Leon Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski, 2003. *Building a discourse-tagged corpus in the framework of rhetorical structure theory*, chapter 5, pages 85–112. Springer Berlin Heidelberg.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 61–66, Baltimore, MD, USA.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.

Ronald Aylmer Fisher. 1932. *Statistical Methods for Research Workers*. Oliver and Boyd, London.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria.

Brigitte Grote, Nils Lenke, and Manfred Stede. 1997. Ma(r)king concessions in English and German. *Discourse Processes*, 24(1):87–118.

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

Gerhard Helbig and Joachim Buscha. 1996. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Langenscheidt.

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING): System Demonstrations*, pages 105–109, Dublin, Ireland.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Sentiment analysis in twitter with lightweight discourse analysis. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1847–1864, Mumbai, India.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, New York, NY, USA.

Gary Patterson and Andrew Kehler. 2013. Predicting the Presence of Discourse Connectives. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 914–923, Seattle, Washington, USA.

Andreas Peldszus and Manfred Stede. 2013a. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 28–30, Marrakech, Morocco.

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar.

Manfred Stede. 2002. DiMLex: A Lexical Approach to Discourse Markers. In Alessandro Lenci and Vittorio Di Tomaso, editors, *Exploring the Lexicon. Theory and Computation*. Edizioni Dell'Orso, Alessandria.

Maite Taboada and María de los Angeles Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, 6(1-3):17–41.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Maite Taboada. 2006. Discourse Markers as Signals (or Not) of Rhetorical Relationsteu. *Journal of Pragmatics*, 38(4):567–592.

Artem Vovk. 2013. Discovery and Analysis of Public Opinions on Controversial Topics in the Educational Domain. Master's thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.

Bonnie Webber, Mark Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18:437–490, 10.

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 13–22, Hannover, Germany.

Florian Wolf and Edward Gibson. 2005. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics*, 31(2):249–288.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2004. Discourse Graphbank. Linguistic Data Consortium.