An Argument-Annotated Corpus of Scientific Publications

Anne Lauscher,[1,2] **Goran Glavaš,**[1] and **Simone Paolo Ponzetto**[1]
[1]Data and Web Science Research Group
University of Mannheim, Germany
[2]Web-based Information Systems and Services
Stuttgart Media University, Germany
{anne, goran, simone}@informatik.uni-mannheim.de
lauscher@hdm-stuttgart.de

Abstract

Argumentation is an essential feature of scientific language. We present an annotation study resulting in a corpus of scientific publications annotated with argumentative components and relations. The argumentative annotations have been added to the existing Dr. Inventor Corpus, already annotated for four other rhetorical aspects. We analyze the annotated argumentative structures and investigate the relations between argumentation and other rhetorical aspects of scientific writing, such as discourse roles and citation contexts.

1 Introduction

With the rapidly growing amount of scientific literature (Bornmann and Mutz, 2015), computational methods for analyzing scientific writing are becoming paramount. To support learning-based models for automated analysis of scientific publications, potentially leading to better understanding of the different rhetorical aspects of scientific language (which we dub *scitorics*), researchers publish manually-annotated corpora. To date, existing manually-annotated scientific corpora already reflect several of these aspects, such as sentential discourse roles (Fisas et al., 2015), *argumentative zones* (Teufel et al., 1999, 2009; Liakata et al., 2010), subjective aspects (Fisas et al., 2016), and citation polarity and purpose (Jochim and Schütze, 2012; Jha et al., 2017; Fisas et al., 2016).

As tools of persuasion (Gilbert, 1976, 1977), scientific publications are abundant with argumentation. Yet, somewhat surprisingly, there is no publicly available corpus of scientific publications (in English), annotated with fine-grained argumentative structures. In order to support comprehensive analyses of rhetorics in scientific text (i.e., *scitorics*), argumentative structure of scientific publications should not be studied in isolation, but rather in relation to other rhetorical aspects, such as the discourse structure. This is why in this work we contribute a new argumentation annotation layer to an existing Dr. Inventor Corpus (Fisas et al., 2016), already annotated for several rhetorical aspects.

Contributions. We propose a general argument annotation scheme for scientific text that can cover various research domains. We next extend the Dr. Inventor corpus (Fisas et al., 2015, 2016) with an annotation layer containing fine-grained argumentative components and relations. Our efforts result in the first argument-annotated corpus of scientific publications (in English), which allows for joint analyses of argumentation and other rhetorical dimensions of scientific writing. We make the argument-annotated corpus publicly available.[1] Finally, we offer an extensive statistical and information-theoretic analysis of the corpus.

2 Related Work

Researchers have offered a plethora of argument annotation schemes and corpora for various domains, including Wikipedia discussions (Biran and Rambow, 2011), on-line debates (e.g., Abbott et al., 2016; Habernal and Gurevych, 2016), e-markets (e.g., Islam, 2007), persuasive essays (Stab and Gurevych, 2017), news editorials (Al Khatib et al., 2016), and law (Wyner et al., 2010). The corpus of Reed et al. (2008) covers multiple domains, including news and political debates.

The work on argumentative annotations in scientific writing is, however, much scarcer. Pioneering annotation efforts of Teufel and Moens (1999a,b); Teufel et al. (1999) focused on discourse-level argumentation (dubbed *argumentative zones*), denoting more the rhetorical structure of the publica-

[1] http://data.dws.informatik.uni-mannheim.de/sci-arg/compiled_corpus.zip

Proceedings of the 5th Workshop on Argument Mining, pages 40–46

tions than fine-grained argumentation, i.e., there are no (1) fine-grained argumentative components (at sub-sentence level) and no (2) relations between components, giving rise to an argumentation graph. Blake (2010) distinguishes between explicit and implicit claims, correlations, comparisons, and observations in biomedical publications. In contrast, we are not interested in how the claim is made, but rather on what are the claims (and what is not a claim) and how they are mutually connected. Green et al. (2014); Green (2014, 2015, 2016) proposed methods for identifying and annotating argumentative structures in scientific publications, but released no publicly available annotated corpus. In the effort most similar to ours, Kirschner et al. (2015) annotated arguments in a corpus of educational research publications. Besides being quite small, this corpus is also written in German.

3 Annotation Scheme

A number of theoretical frameworks of argumentation have been proposed (Walton et al., 2008; Anscombre and Ducrot, 1983, *inter alia*).[2] Among the most widely used is the model of Toulmin (2003), from which we start in this work as well, because of its relative simplicity and adoption in artificial intelligence and argument mining (Bench-Capon, 1998; Verheij, 2005; Kirschner et al., 2015). The Toulmin model, originally developed for the legal domain, recognizes six types of argumentative components: *claim*, *data*, *warrant*, *backing*, *qualifier*, and *rebuttal*.

We conducted a preliminary annotation study using the Toulmin model with two expert annotators on a small corpus subset. Annotators did not identify any *warrant*, *backing*, *qualifier*, nor *rebuttal* components. The annotators also pointed to the interlinked argumentative structure of publications in which *claim* were often used as ground for (supporting or conflicting) another claim. Not foreseen by the Toulmin model, we realized that the relations between argumentative components can be of different nature. Finally, the annotators recognized two distinct claim types: those presented as common knowledge (or state of the art) in the research area and those relating to authors' own research.

Following the above observations from the preliminary annotation, we simplify the annotation scheme by removing the non-observed component types. Our final annotation scheme has the following types of argumentative component:

(1) *Own Claim* is an argumentative statement that closely relates to the authors' own work, e.g.:

> *"Furthermore, we show that by simply changing the initialization and target velocity, the same optimization procedure leads to running controllers."*

(2) *Background Claim* is an argumentative statement relating to the background of authors' work, e.g., about related work or common practices in the respective research field, e.g.:

> *"Despite the efforts, accurate modeling of human motion remains a challenging tasks."*

(3) *Data* component represents a fact that serves as evidence for or against a claim. Note that references or (factual) examples can also serve as data, e.g.:

> *"[...], due to memory and graphics hardware constraints nearly all video game character animation is still done using traditional SSD."*

We follow Bench-Capon (1998) and allow for links between the arguments. We introduce three different relations types, similar to Dung (1995).

(1) A *Supports* relation holds between components a and b if the assumed veracity of b increases with the veracity of a;
(2) A *Contradicts* relation holds between components a and b if the assumed veracity of b decreases with the veracity of a;
(3) The *Semantically Same* relation is annotated between two mentions of effectively the same claim or data component. This relation can be seen as *argument coreference*, analogous to entity (Lee et al., 2011, 2017) and event coreference (Glavaš and Šnajder, 2013; Lu and Ng, 2018).

It is important to emphasize that we do not bind the spans of our argumentative components to sentence boundaries, but rather allow for argumentative components of arbitrary span lengths, ranging from a single token to multiple sentences.

4 Annotation study

Dataset. Believing that argumentation needs to be studied in combination with other rhetorical aspects of scientific writing, we enriched the existing Dr. Inventor corpus (Fisas et al., 2015, 2016), consisting of 40 publications from computer graphics, with argumentative information. The Dr. Inventor

[2] For an extensive overview, we refer the reader to (Bentahar et al., 2010)

Annotation Layer	Labels	%
Discourse Role	*Background*	20
	Challenge	5
	Approach	57
	Outcome	16
	Future Work	2
Citation Purpose	*Criticism*	23
	Comparison	9
	Use	11
	Substantiation	1
	Basis	5
	Neutral	53
Subjective Aspect	*Advantage*	33
	Disadvantage	16
	Adv.-Disadv.	3
	Disadv.-Adv.	1
	Novelty	13
	Common Practice	32
	Limitation	2
Summarization Relevance	*Totally irrelevant*	66
	Should not appear	6
	May appear	14
	Relevant	6
	Very relevant	8

Table 1: Annotation layers of the Dr. Inventor Corpus with label distributions.

corpus has four layers of rhetorical annotations: (1) discourse roles, (2) citation purposes with associated citation contexts, (3) judgments of subjective aspects, and (4) annotations of sentence relevance for a summary. Table 1 summarizes the different annotation layers and their label distributions.

Annotation Process. We hired one expert[3] and three non-expert annotators[4] for our annotation study. We trained the annotators in a calibration phase, consisting of five iterations, in each of which all annotators annotated one publication. After each iteration we computed the inter-annotator agreement (IAA), discussed the disagreements, and, if needed, adjourned the annotation guidelines.[5] We measured the IAA in terms of the F_1-measure because (1) it is easily interpretable and straight-forward to compute and (2) it can account for spans of varying length, allowing for computing relaxed agreements in terms of partial overlaps.[6] The evolution of IAA over the five calibration iterations is depicted in Figure 1, in two variants: (1) A *strict* version in which components have to match exactly in span and type and relations have to match exactly in both components, direction and type of the link and (2) a *relaxed* version in which components only have to match in type and overlap in span (by at least half of the length of the shorter of them). Expectedly, we observe higher agreements with more calibration. The agreement on argumentative relations is 23% lower than on the components, which we think is due to the high ambiguity of argumentation structures, as previously noted by Stab et al. (2014). That is, given an argumentative text with pre-identified argumentative components, there are often multiple valid interpretations of an argumentative relation between them, i.e., it is "[...] hard or even impossible to identify one correct interpretation" (Stab et al., 2014). Additionally, disagreements in component identification are propagated to relations as well, since the agreement on a relation implies the agreement on annotated components at both ends of the relation.

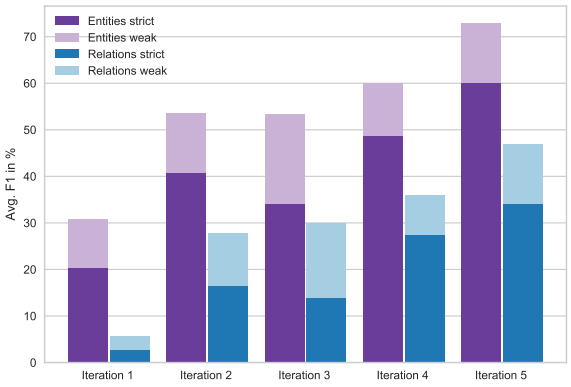

Figure 1: IAA evolution over the five calibration phases (*purple* for argumentative components; *blue* for relations; *dark* for the *strict* agreements; *light* for the *relaxed* agreements).

5 Corpus Analysis

We first study the argumentation layer we annotated in isolation. Afterwards, we focus on the interrelations with other rhetorical annotation layers.

Analysis of Argumentation Annotations. Table 2 lists the number of components and relations in total and on average per publication. The number of *own claims* roughly doubles the amount of *background claims*, as the corpus consists only of original research papers, in which the authors mainly emphasize their own contributions. Interest-

[3] A researcher in computational linguistics, not in computer graphics.

[4] Humanities and social sciences scholars.

[5] http://data.dws.informatik.uni-mannheim.de/sci-arg/annotation_guidelines.pdf

[6] Note that the chance-corrected measures, e.g., Cohen's Kappa, approach F_1-measure when the number of negative instances grows (Hripcsak and Rothschild, 2005).

Category	Label	Total	Per Publication
Component	*Background claim*	2,751	68.8 ± 25.2
	Own claim	5,445	136.1 ± 46.0
	Data	4,093	102.3 ± 32.1
Relation	*Supports*	5,790	144.8 ± 43.1
	Contradicts	696	17.4 ± 9.1
	Semantically same	44	1.1 ± 1.81

Table 2: Total and per-publication distributions of labels of argumentative components and relations in the extended Dr. Inventor Corpus.

Label	Min	Max	Avg (μ)	Std (σ)
Background claim	5	340	87.46	43.74
Own claim	3	500	85.70	44.03
Data	1	244	25.80	27.59

Table 3: Statistics on length of argumentative components (in number of characters) in the extended Dr. Inventor Corpus.

ingly, there are only half as many *data* components as claims. We can see two reasons for this – first, not all claims are supported and secondly, claims can be supported by other claims. There are many more *supports* than *contradicts* relations. This is intuitive, as authors mainly argue by providing *supporting* evidence for their own claims.

Table 3 shows the statistics on length of argumentative components. While the *background claims* and *own claims* are on average of similar length (85 and 87 characters, respectively), they are much longer than *data* components (average of 25 characters). This is intuitive given the domain of the corpus, as facts in computer science often require less explanation than claims. For example, we noticed that authors often refer to tables and figures as evidence for their claims. Similarly, when claiming weaknesses or strengths of related work, authors commonly provide references as evidence.

The argumentative structure of an individual publication corresponds to a forest of directed acyclic graphs (DAG) with annotated argumentative components as nodes and argumentative relations as edges. Thus, to obtain further insight into structural properties of argumentation in scientific publications, in Table 4 we provide graph-based measures like the number of connected components (i.e., subgraphs), the diameter, and the number of standalone claims (i.e., nodes without incoming or outgoing edges) and unsupported claims (i.e., nodes with no incoming *supports* edges). Our

Criterion	Min	Max	Avg (μ)	Std (σ)
Diameter	2	5	3.05	0.71
Max In-Degree	3	11	6.33	1.97
# standalone claims	27	127	63.00	21.40
# unsupp. claims	39	180	94.38	29.14
# unconn. subgraphs	78	231	147.23	35.78
# comp. per subgraph	1	17	2.09	1.5

Table 4: Graph-based analysis of the argumentative structures identified in the extended Dr. Inventor Corpus (per publication).

annotators identified an average of 141 connected component per publication, with an average diameter of 3. This indicates that either authors write very short argumentative chains or that our annotators had difficulties noticing long-range argumentative dependencies.

On the one hand, there are at least 27 standalone claims in each publication, that is claims, that are not connected with any other components. On the other hand, the maximum in-degree of a claim in a publication, on average, is 6, indicating that there are claims for which a lot of evidence is given. Intuitively, the claims for which more evidence is given should be more prominent. We next run PageRank (Page et al., 1999) on argumentation graphs of individual publications to identify most prominent claims. We list a couple of examples of claims with highest PageRank scores in Table 5. Somewhat unexpectedly, in 30 out of 40 publications in the dataset the highest ranked claim was a *background claim*. This suggests that in computer graphics authors emphasize more research gaps and motivation for their work than they justify its impact (for which empirical results often suffice).

Links to Other Rhetorical Aspects. We next investigate the interdependencies between the newly added argumentative annotations and the existing rhetorical annotations of the *Dr. Inventor Corpus*. An inspection of dependencies between different annotation layers in the corpus may indicate the usefulness of computational approaches that aim to exploit such interrelations. E.g., Bjerva (2017) recently showed that the measure of mutual information strongly correlates with performance gains obtained by multi-task learning models.

In this work, we employ the measure of normalized mutual information (NMI) (Strehl and Ghosh, 2003) to assess the amount of information shared between the five annotation layers. NMI is a variant of mutual information scaled to the interval $[0, 1]$

Type	Pub.	Claim with maximal PageRank score
background claim	A13	'physical validity is often sacrificed for performance'
	A21	'a tremendous variety of materials exhibit this type of behavior'
own claim	A39	'the solution to the problem of asymmetry is to modify the CG method so that it can operate on equation (15), while procedurally applying the constraints inherent in the matrix W at each iteration'

Table 5: Claims with maximum PageRank score in a publication.

	AC	DR	SA	SR
AC	–	–	–	–
DR	**0.22**	–	–	–
SA	0.08	0.11	–	–
SR	0.04	0.10	0.13	–
CC	**0.18**	0.10	0.04	0.01

Table 6: Normalized mutual information between different annotation layers.

through normalization with the entropy of each of the two label sets. In Table 6 we show the NMI scores for all pairs of annotations layers: argument components (AC), discourse roles (DR), citation contexts (CC), subjective aspects (SA), and summary relevances (SR). The strongest association is found between argumentative components (AC) and discourse roles (DR). Looking at the labels of these two annotation layers, this seems plausible – *background claim* (AC) is likely to appear in a sentence of discourse role *background* (DR). Similarly, *own claims* more frequently appear in sections describing the *outcomes* of the work. To confirm this intuition, we computed co-occurrence matrices for pairs of label sets – indeed, the AC label *own claim* most frequently appears together with the discourse role *approach* and *outcome*, and the *background claim* with discourse roles *background* and *challenge*. Consider the following sentence:

> *"With the help of modeling tools or capture devices, complicated 3D character models are widely used in the fields of entertainment, virtual reality, medicine, etc."*

It contains a general claim about the research area (i.e., it is a *background claim*) and it also offers *background* information in terms of the discourse role. A similar set of intuitive label alignments justifies the higher NMI score between argumentative components (AC) and citation contexts (CC): *citation contexts* often appear in sentences with a *background claim*. Again, this is not surprising, as authors typically reference other publications and in order to motivate their work:

> *"An improvement based on addition of auxiliary joints has been also proposed in [Weber 2000]. Although this reduces the artifacts, the skin to joints relationship must be re-designed after joint addition."*

In the above example, the wave-underlined text, i.e. the citation, serves as the *data* for the underlined text which is the *background claim* stating a research gap in the referenced work. At the same time, the underlined text can be seen as the *citation context* with the reference as target.

6 Conclusion

We presented an annotation scheme for argumentation analysis in scientific publications. We annotated the *Dr. Inventor Corpus* (Fisas et al., 2015, 2016) with an argumentation layer. The resulting corpus, which is, to the best of our knowledge, the first argument-annotated corpus of scientific publications in English, enables (1) computational analysis of argumentation in scientific writing and (2) integrated analysis of argumentation and other rhetorical aspects of scientific text. We further provided corpus statistics and graph-based analysis of the argumentative structure of the annotated publications. Finally, we analyzed the dependencies between different rhetorical aspects, which can inform computational models aiming to jointly address multiple aspects of scientific discourse. In the future, we plan to extend the corpus with publications from other domains and develop computational models for the integrated analysis of scientific writing.

Acknowledgments

This research was partly funded by the German Research Foundation (DFG), grant number EC 477/5-1 (LOC-DB). We thank our annotators for their very dedicated annotation effort.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4445–4452, Portorož, Slowenia. European Language Resources Association.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'argumentation dans la langue*. Editions Mardaga.

Trevor JM Bench-Capon. 1998. Specification and implementation of toulmin dialogue game. In *Proceedings of the 11th Conference on Legal Knowledge Based Systems*, pages 5–20, Groningen, Netherlands. Foundation for Legal Knowledge Based Systems.

Jamal Bentahar, Bernard Moulin, and Micheline Blanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Fifth IEEE International Conference on Semantic Computing*, pages 162–168, Palo Alto, CA, USA. IEEE.

Johannes Bjerva. 2017. Will my auxiliary tagging task help? estimating auxiliary tasks effectivity in multi-task learning. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 216–220, Gothenburg, Sweden. Association for Computational Linguistics.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, CO, USA. Association for Computational Linguistics.

G Nigel Gilbert. 1976. The transformation of research findings into scientific knowledge. *Social Studies of Science*, 6(3-4):281–306.

G Nigel Gilbert. 1977. Referencing as persuasion. *Social Studies of Science*, 7(1):113–122.

Goran Glavaš and Jan Šnajder. 2013. Exploring coreference uncertainty of generically extracted event mentions. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 408–422. Springer.

Nancy Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland. Association for Computational Linguistics.

Nancy Green. 2015. Annotating evidence-based argumentation in biomedical text. In *2015 IEEE International Conference on Bioinformatics and Biomedicine*, pages 922–929, Washington, D.C., USA. IEEE.

Nancy Green. 2016. Implementing argumentation schemes as logic programs. In *The 16th Workshop on Computational Models of Natural Argument*, volume 30, New York, USA. CEUR-WS.

Nancy Green, E Cabrio, S Villata, and A Wyner. 2014. Argumentation for scientific claims in a biomedical research article. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 21–25, Forl-Cesena, Italy. CEUR-WS.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1122, Berlin, Germany. Association for Computational Linguistics.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Khandaker Shahidul Islam. 2007. An Approach to Argumentation Context Mining from Dialogue History in an e-Market Scenario. In *Proceedings of the 2Nd International Workshop on Integrating Artificial Intelligence and Data Mining - Volume 84*, AIDM '07, pages 73–81, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R. Radev. 2017. NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, 23(1):93–130.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1343–1358, Mumbai, India. The COLING 2012 Organizing Committee.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11, Denver, CO, USA. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin R Batchelor. 2010. Corpora for the Conceptualisation and Zoning of Scientific Papers. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Jing Lu and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *IJCAI*, pages 5479–5486.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2613–2618, Marrakesh, Marocco. European Language Resources Association.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 21–25.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 1999a. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In *Advances in automatic Text Summarization*, pages 155–171, Cambridge, MA, USA. MIT Press.

Simone Teufel and Marc Moens. 1999b. Discourse-level argumentation in scientific articles: Human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging, Workshop*, Maryland, MA, USA. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, pages 1493–1502, Edinburgh, Scotland. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*, updated edition. Cambridge University Press.

Bart Verheij. 2005. Evaluating Arguments Based on Toulmins Scheme. *Argumentation*, 19(3):347–371.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to Text Mining Arguments from Legal Cases. In *Semantic Processing of Legal Texts*, pages 60–79. Springer Berlin Heidelberg.