AMResources: Cataloging Argument Mining Datasets

Dexter Williams[1], Shiwei Liu[4], Manfred Stede[2], Henning Wachsmuth[3], Jodi Schneider[1]

[1]University of Wisconsin-Madison, [2]Potsdam University, [3]Leibniz University Hannover, [4]University of Illinois Urbana-Champaign

Correspondence: djwilliams22@wisc.edu and jschneider@pobox.com/jodi@wisc.edu

Abstract

Annotated datasets are essential for developing and evaluating argument mining systems, yet information about argument mining datasets remains scattered across papers, repositories, and task-specific surveys. To address this, we introduce *AMResources* (http://purl.archive.org/amresources), an online catalog that organizes argument mining datasets by task, and captures relationships among datasets, releases, and papers. We draw particular attention to relationships such as re-annotation and dataset extension. To curate dataset information into a consistent and provenance-aware structure, *AMResources* links datasets to canonical papers. For each dataset release, *AMResources* records standardized metadata such as language, genre, unit type and unit count, annotator characteristics, agreement reporting, and accessibility. We argue that such structured dataset documentation remains critical in the era of large language models, where annotated datasets increasingly serve as high-quality evaluation benchmarks and where tracing dataset provenance and annotation layers is necessary for systematic comparisons across tasks.

1 Introduction

Datasets are an essential resource for argument mining tasks, supporting both the training of computational models and the evaluation of automated methods. However, a number of factors (theoretical and practical) affect the cataloging of argument mining datasets. Argument mining datasets may be annotated for a number of different tasks and subsequently annotated for a different purpose (Stab and Gurevych, 2017a; Goffredo et al., 2023). Annotated datasets differ in the quality of annotators used, and may or may not report agreement. Datasets vary in their availability, size, language, and format. A dataset may also stand in an important relationship to other datasets, through the inheritance of previous layers of annotation or subsets of units.

Despite these challenges, argument mining datasets have been cataloged and discussed by the community. Habernal and Gurevych (2017) provide a table of 23 previous works on annotating argumentation, documenting the properties of created datasets. A chapter of *Argumentation Mining* (Stede and Schneider, 2018) examines several annotation schemes and catalogs 10 relevant datasets. Janier and Saint-Dizier (2019) includes a chapter on argument annotation, covering annotation schemes, tools, and 8 relevant datasets. Lawrence and Reed (2019) discusses a number of datasets, focusing on the challenges of producing and annotating datasets for argument mining.

A number of online sites have been created to make argument mining-related resources more findable, although their scope and maintenance varies. Persiani et al. (2024) provides a website[1] containing an ongoing survey of tools useful for computational argumentation; the site is frequently updated and accepts content requests from the community through a GitHub repository.[2] Guerraoui et al. (2023) surveys NLP feedback systems in argumentation; a supporting website[3] categorizes works on datasets, tooling, and computational models into four sections (richness, visualization, interactivity, and personalization).

A few sites catalog argument mining datasets. The Ubiquitous Knowledge Processing (UKP) Lab at TU Darmstadt[4] and the Webis Group[5] publish datasets used in their argument mining research.

[1]https://people.cs.umu.se/~tkampik/argtools/page/index.html

[2]https://github.com/TimKam/fantastic-arg-tools

[3]https://kmilia.github.io/teach_me_how_to_argue/

[4]https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/1359

[5]https://webis.de/data.html

AIFdb (Lawrence and Reed, 2014; Lawrence et al., 2016) is a large community-sourced database of argument mining datasets.[6] Datasets in *AIFdb* are stored in the Argument Interchange Format (Chesñevar et al., 2006), which enables interoperability with argument mapping tools such as *OVA+* (Janier et al., 2014). However, *AIFdb* does not enforce documentation of dataset properties. This makes it challenging to use *AIFdb* to search for datasets for a specific argument mining task. In contrast, the documentation of dataset properties is more evident in the website introduced by Romberg et al. (2025) to accompany their survey of datasets for automatic argument quality assessment.[7] Their website categorizes argument quality datasets according to a taxonomy, and lists a number of properties of each dataset such as annotator agreement, annotator type, source, and whether the dataset is an extension of another dataset.

In this paper we introduce an online catalog of argument mining datasets: *AMResources*.[8] In contrast to other sites, *AMResources* emphasizes the suitability of datasets for specific argument mining tasks, and makes dataset links findable. *AMResources* catalogs argument mining datasets across a plethora of argument mining tasks through a linked structure of *datasets*, *releases*, *papers*, and *annotations*. Our work is inspired by the catalog of Romberg et al. (2025). During the creation of Stede et al. (2026), the second edition of (Stede and Schneider, 2018), it became apparent that the rapid growth of argument mining datasets could not be covered in a single chapter; this led to the development of our online catalog as a maintainable resource for discovering argument mining datasets. The initial release of *AMResources* contains the 39 datasets described in the 'Selected Corpora' appendix of Stede et al. (2026).

2 Connecting Datasets, Papers, and Annotations

We consider a set of units (documents, tweets, etc.) to be a *dataset*. A dataset has a number of properties such as a *domain*, *unit count*, and *unit type*. The first publication of a *dataset* is the *source dataset*. Any newer dataset which was published after the source dataset, with units that are a subset of the units of the source dataset, is a *release* of the source dataset.

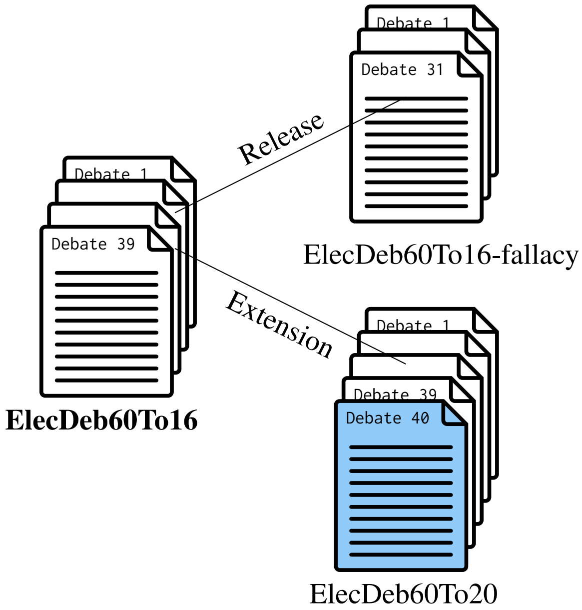

Figure 1: Release and extension relationships between the ElecDeb series of datasets. ElecDeb60To16 is a source dataset. ElecDeb60To16-fallacy is a release of ElecDeb60To16 as it consists of a subset of debates from ElecDeb60To16. ElecDeb60To20 extends ElecDeb60To16 with a new debate transcript.

Releases serve an important function in capturing inheritance relationships between datasets. The units within one dataset might just be a subset of the units from another dataset. Frequently, datasets are annotated multiple times, with each annotated dataset being given a new name. For example the ElecDeb60To16 dataset (Haddadan et al., 2019), a dataset of political transcripts from 1960 to 2016, was subsequently annotated and released as the ElecDeb60To16-fallacy dataset (Goffredo et al., 2022). Thus, ElecDeb60To16 is the source dataset, with a later release in ElecDeb60To16-fallacy.

It is also possible for a newer dataset to include a subset of the units of a source dataset while also adding new units. In this case we do not call the newer dataset a release of the source dataset, but rather an *extension* of the source dataset. For example the ElecDeb60To20 dataset (Goffredo et al., 2023) contains an additional four years of political transcripts from 2016 to 2020 as well as the original political transcripts in ElecDeb60To16. Therefore we call ElecDeb60To20 an *extension* of ElecDeb60To16. Relationships between the three ElecDeb datasets can be seen in Figure 1.

Our approach allows us to capture the evolution

[6] https://www.aifdb.org/search
[7] https://goofy-grouse-1da.notion.site/Database-e3e5886191ef472aaffb47fec0daea92
[8] http://purl.archive.org/amresources

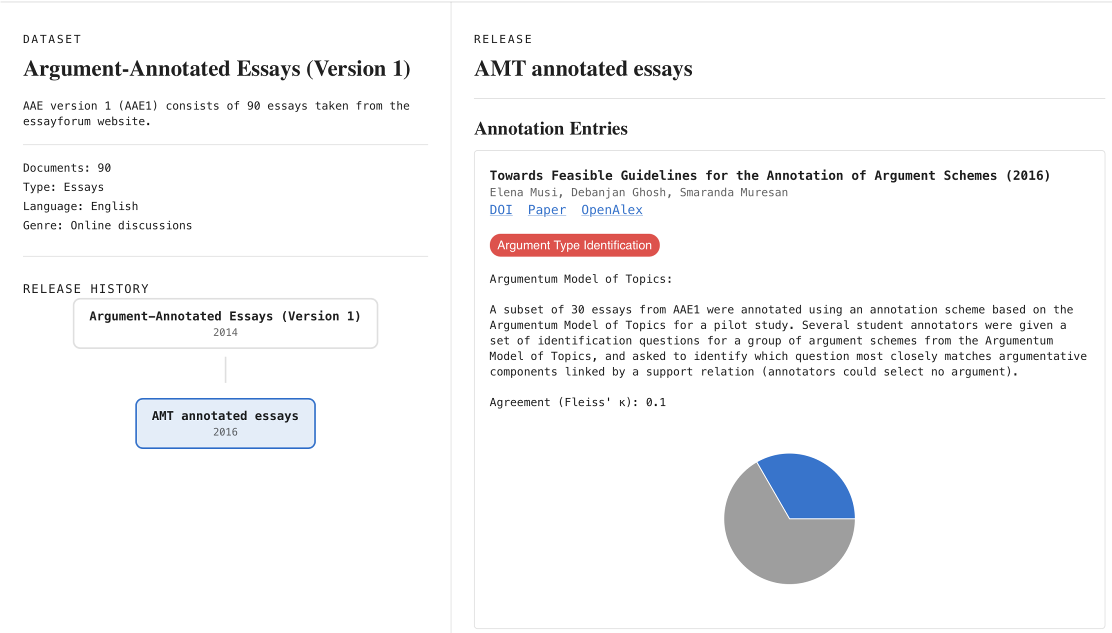

Figure 2: Example view from *AMResources*. The user has selected details on AMT annotated essays (Musi et al., 2016) which is a release of the Argument Annotated Essays Corpus (Stab and Gurevych, 2014). The left pane displays other releases for the dataset. The right pane shows details on the release, including a pie chart showing how much of the original dataset was annotated in the release.

of properties of argument mining datasets, instead of presenting a flat table of datasets. Datasets that have been annotated multiple times in different releases are particularly valuable to identify. Multi-layered argument annotations on a dataset can be used to identify dependencies between different argumentative relations (Chistova, 2023). Some argument mining datasets have been annotated repeatedly, for instance the Argument-Annotated Essays Corpus has been annotated in at least eight different papers (Carlile et al., 2018; Stab and Gurevych, 2014, 2016, 2017a,b; Musi et al., 2016; Schaefer et al., 2023; Marro et al., 2022).

Finally, a *paper* is an academic work that either introduces a source dataset or annotates an existing dataset to produce a new release of that dataset. An *annotation* is carried out by a paper on a dataset, to support a particular argument mining task (see Section 3). *Datasets*, *releases*, *papers*, and *annotations* form the four linked objects of *AMResources*.

3 Argument Mining Tasks

Beyond the surface-level properties of datasets, we aim to categorize argument mining datasets by their potential use. In particular, we map each dataset to the set of argument mining tasks that it enables studying. We differentiate between two broad categories of argument mining tasks based on their use of argumentative discourse units (ADUs): *ADU-concerned* and *ADU-agnostic*.

ADU-Concerned Tasks This category covers tasks that, in some way, deal with the collection, classification, and linking of ADUs. Our selection of ADU-concerned tasks is taken from Stede et al. (2026, Ch. 4) and forms a classic argument mining pipeline. The description of ADU-concerned tasks is shown in Table 1.

ADU-Agnostic Tasks These tasks do not necessarily involve ADUs but undebatably fall under the notion of argument mining in the broader sense (see Table 2). For instance, argument quality assessment might not involve annotating argument structure, and instead assigning binary scores of quality to argumentative text (Toledo et al., 2019). We distinguish between minimal and maximal argument quality assessment, a distinction taken from Wachsmuth et al. (2024). Another ADU-agnostic task is to mine the full type/form/scheme of an argument from text (e.g., Schneider et al., 2013); we call this argument type identification.

4 AMResources: A Website for Argument Mining Datasets

We host our collection of datasets on a website called *AMResources*, deployed as a *GitHub Pages*

Task	Description
Component Segmentation	Distinguishes argumentative from non-argumentative text portions and demarcates individual argument components
Component Type Classification	Determines the types of given argument components, using a coarse-grained (e.g., claim, premise) or a more fine-grained (e.g., policy claim, fact claim, value claim) type inventory
Relation Identification	Detects the presence of some (indeterminate) argument relation between two argument components
Relation Type Classification	Determines the types of given relations (e.g., support, attack)

Table 1: ADU-concerned tasks: The four tasks that deal with the collection, classification, and linking of ADUs, from Stede et al. (2026, Ch. 4).

Task	Description
Maximal Quality Assessment	Assesses the quality of an argument based on a theory that describes what an argument should *ideally* be
Minimal Quality Assessment	Assesses the quality of an argument based on a theory that describes what an argument should *avoid* being
Type Identification	Identifying the type/form of the whole argument

Table 2: Three examples of ADU-agnostic tasks, from Wachsmuth et al. (2024) and (Schneider et al., 2013).

site.[9] The repository for the site is public, and we encourage contributions of new datasets, and include a form for requests for new datasets to be added. We store all covered dataset metadata in two JSON files (one for dataset metadata and one for paper metadata) which can be downloaded through the GitHub repository.

AMResources presents releases in a searchable table view (release, dataset, language, genre, unit type, unit count, and task tags). The catalog can be exported as a CSV file. Selecting a row opens a detail view (see Figure 2) with dataset metadata, a release history visualization, and structured annotation summaries linked to papers. Where possible we provide a DOI and an OpenAlex link for papers. Each annotation entry contains a detailed summary, and records task labels, annotator type, agreement reporting, accessibility, and a release link.

5 The Future Role of Argument Mining Datasets

Argument mining datasets have been, and continue to be, an important artifact of argument mining research. However, it is prudent to reflect on the role that argument mining datasets now serve in light of advancements in large language models (LLMs). The rise in the use of LLMs reflects a shift away from classical supervised learning methods which relied on annotated datasets, towards methods which involve little to no need for manually annotated data for training.

Despite a decreasing relevance to model training, argument mining datasets continue to be important for evaluation. Several argument mining tasks (such as ADU segmentation) are not straightforward to model with LLMs, leading to a demand for LLM-oriented argument mining benchmarks (Dhole et al., 2025; Gemechu et al., 2024; Gurjar et al., 2025; Ajjour et al., 2026).

Additionally, annotated datasets in argument mining also have a unique role in comparison to other fields. Argument mining has an unusually strong theoretical foundation compared to many other NLP tasks. This heavy emphasis on theory is manifested in the annotations, and it has the potential to inform NLP methods that go beyond pure statistics (Lauscher et al., 2022; Wachsmuth et al., 2024). In particular, competing theoretical models of argument underlie argument mining. These competing theoretical models make their own assumptions and simplifications (Cardoso et al., 2023). Datasets with multiple layers of annotation provide an opportunity to study dependencies between argument mining tasks and argument models. In this respect, annotated argument mining datasets can help us to inform our theoretical models of arguments, and how these models interface with text.

[9] http://purl.archive.org/amresources

6 Conclusion

We presented AMResources, a web catalog for argument mining datasets that represents provenance through explicit links among datasets, releases, and papers. We argue that such structured dataset documentation remains critical in the era of large language models, where annotated datasets increasingly serve as high-quality evaluation benchmarks and where tracing dataset provenance and annotation layers is necessary for systematic comparisons across tasks. Going forward, we will expand coverage through community contributions in order to keep *AMResources* up-to-date over time.

Acknowledgments

Dexter Williams was supported by University of Wisconsin-Madison startup funds to Jodi Schneider as well as her University of Illinois Campus Research Board RB25118: Support for a Second Edition of the Argumentation Mining monograph, which also supported work from Shiwei Liu. Some support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. Jodi Schneider was partially funded as the 2024-2025 Perrin Moorhead Grayson and Bruns Grayson Fellow at Harvard University's Radcliffe Institute for Advanced Study. Henning Wachsmuth is partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/3 2026 – 438445824.

CRediT statement

Dexter Williams: Writing - Original Draft, Data Curation, Software, Project Administration
Shiwei Liu: Data Curation, Writing - Review & Editing
Manfred Stede: Conceptualization, Writing - Review & Editing
Henning Wachsmuth: Conceptualization, Writing - Review & Editing
Jodi Schneider: Conceptualization, Supervision, Funding Acquisition, Writing - Review & Editing
Markus Rutledge: Data Curation

References

Yamen Ajjour, Carlotta Quensel, Nedim Lipka, and Henning Wachsmuth. 2026. ArgBench: Benchmarking LLMs on computational argumentation tasks. *Preprint*, arXiv:2604.17366.

Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins. 2023. Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an Argument Interchange Format. *The Knowledge Engineering Review*, 21(4):293–316.

Elena Chistova. 2023. End-to-end argument mining over varying rhetorical structures. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3376–3391, Toronto, Canada. Association for Computational Linguistics.

Kaustubh Dhole, Kai Shu, and Eugene Agichtein. 2025. ConQRet: A new benchmark for fine-grained automatic evaluation of retrieval augmented computational argumentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5687–5713, Albuquerque, New Mexico. Association for Computational Linguistics.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Main Track*, volume 5, pages 4143–4149, Vienna, Austria. International Joint Conferences on Artificial Intelligence.

Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on NLP feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*, pages 19–34, Singapore. Association for Computational Linguistics.

Omkar Gurjar, Agam Goyal, and Eshwar Chandrasekharan. 2025. ArgCMV: An argument summarization benchmark for the LLM-era. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21870–21883, Suzhou, China. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: An argument analysis interface. In *Computational Models of Argument*, pages 463–464. IOS Press.

Mathilde Janier and Patrick Saint-Dizier. 2019. *Argument mining: Linguistic foundations*. Wiley-ISTE, Hoboken.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence, Mathilde Janier, and Chris Reed. 2016. Working with open argument corpora. In *Argumentation and Reasoned Action: Proceedings of the First European Conference on Argumentation*, pages 367–380, Lisbon, Portugal. College Publications.

John Lawrence and Chris Reed. 2014. AIFdb corpora. In *Computational Models of Argument*, pages 465–466. IOS Press.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2016. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining*, pages 82–93, Berlin, Germany. Association for Computational Linguistics.

Michele Persiani, Esteban Guerrero, Andreas Brännström, Kaan Kilic, and Timotheus Kampik. 2024. Fantastic argumentation tools and where to find them. In *Proceedings of the Fifth International Workshop on Systems and Algorithms for Formal Argumentation co-located with 10th International Conference on Computational Models of Argument, Hagen, Germany, September 17th, 2024*, CEUR Workshop Proceedings, pages 56–68. CEUR-WS.org.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2023. Towards fine-grained argumentation strategy analysis in persuasive essays. In *Proceedings of the 10th Workshop on Argument Mining*, pages 76–88, Singapore. Association for Computational Linguistics.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work*, CSCW '13, pages 1069–1080, San Antonio, Texas, USA. Association for Computing Machinery.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining*, pages 113–118, Berlin, Germany. Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Manfred Stede, Jodi Schneider, and Henning Wachsmuth. 2026. *Argument mining*, second edition. Springer Nature. Forthcoming.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 1519–1538, Torino, Italia. ELRA and ICCL.