Argotario: Computational Argumentation Meets Serious Games

Ivan Habernal, Raffael Hannemann, Christian Pollak,
Christopher Klamm, Patrick Pauli, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

An important skill in critical thinking and argumentation is the ability to spot and recognize *fallacies*. Fallacious arguments, omnipresent in argumentative discourse, can be deceptive, manipulative, or simply leading to 'wrong moves' in a discussion. Despite their importance, argumentation scholars and NLP researchers with focus on argumentation quality have not yet investigated fallacies empirically. The nonexistence of resources dealing with fallacious argumentation calls for scalable approaches to data acquisition and annotation, for which the *serious games* methodology offers an appealing, yet unexplored, alternative. We present *Argotario*, a serious game that deals with fallacies in everyday argumentation. *Argotario* is a multilingual, open-source, platform-independent application with strong educational aspects, accessible at www.argotario.net.

1 Introduction

Argumentation in natural language has been gaining much interest in the NLP community in recent years. While understanding the structure of an argument is the predominant task of argument mining/computational argumentation (Mochales and Moens, 2011; Stab and Gurevych, 2014; Habernal and Gurevych, 2017), a parallel strand of research tries to assess qualitative properties of arguments (Habernal and Gurevych, 2016b; Stab and Gurevych, 2017). Yet the gap between theories and everyday argumentation, in understanding what 'argument quality' actually is, remains an open research question (Wachsmuth et al., 2017; Habernal and Gurevych, 2016a).

Argumentation theories and critical thinking textbooks, however, offer an alternative view on quality of arguments, namely the notion of *fallacies*: prototypical argument schemes or types that pretend to be correct and valid arguments but suffer logically, emotionally, or rhetorically (Hamblin, 1970). Although this topic was first brought up by Aristotle already some 2,300 years ago, contemporary research on fallacies still does not provide a unifying view and clashes even in the fundamental questions (Boudry et al., 2015; Paglieri, 2016). Nevertheless, there seem to be several types of fallacies, such as *argument ad hominem*,[1] various emotional *appeals*, rhetorical moves of the *red herring*,[2] or *hasty generalization* that are, unfortunately, widely spread in our everyday argumentative discourse. Their powerful and sometimes detrimental impact was revealed in a few manual analyses (Sahlane, 2012; Nieminen and Mustonen, 2014). To the best of our knowledge, there is neither any NLP research dealing with fallacies, nor any resources that would allow for empirical investigation of that matter.

The lack of fallacy-annotated linguistic resources and thus the need for creating and labeling a new dataset from scratch motivated us to investigate *serious games* (also *games with a purpose*)—a scenario in which a task is *gamified* and users (players) enjoy playing a game without thinking much of the burden of annotations (von Ahn and Dabbish, 2008; Mayer et al., 2014). Serious games have been successful in NLP tasks that can be easily represented by images (Jurgens and Navigli, 2014; Kazemzadeh et al., 2014) or that can be simplified to assessing a single word or a pair of propositions (Nevěřilová, 2014; Poesio et al., 2013). More complex tasks such as argument understanding, reasoning, or composing pose several

[1] Attacking the opponent instead of her argument
[2] Distracting to irrelevant issues

Proceedings of the 2017 EMNLP System Demonstrations, pages 7–12
Copenhagen, Denmark, September 7–11, 2017.

design challenges centered around the key question: how to make data creation and annotation efforts fun and entertaining in the first place.

To tackle this open research challenge, we created *Argotario*—an online serious game for acquiring a dataset with fallacious argumentation. The main research **contributions** and features of *Argotario* include:

- Gamification of the fallacy recognition task including player vs. player interaction
- Learning by playing and educational aspects
- Full in-game data creation and annotation, all data are under open license
- Automatic gold label and quality estimation based solely on the crowd
- Multilingual, platform independent, open-source, modular, with native look-and-feel on smartphones

2 Background and Related Work

Fallacies have been an active topic in argumentation theory research in the past several decades. While Aristotle's legacy was still noticeable in the twentieth century, a 'fresh' look by Hamblin (1970) showed that the concept of fallacies as arguments 'that *seem to be* valid but are *not* so' deserves to be put under scrutiny.[3] Theories about fallacies evolved into various categorizations and treatments, ranging from rather practical education-oriented approaches (Tindale, 2007; Schiappa and Nordin, 2013) to rhetorical ones in informal logic (Walton, 1995) or pragma-dialectic (Van Eemeren and Grootendorst, 1987). For a historical overview of fallacies see, e.g., (Hansen, 2015).

Surprisingly, the vast majority of current works on fallacies, and especially textbooks, present only toy examples that one is unlikely to encounter in real life (Boudry et al., 2015, p. 432). The distinction between fallacies and acceptable inference is fuzzy and theories do not offer any practical guidance: fully-fledged fallacies are harder to find in real life than is commonly assumed (Boudry et al., 2015). To this account, analysis of fallacies in actual argumentative discourse has been rather limited in scope and size. Nieminen and Mustonen (2014) examined fallacies found in articles supporting creationism. Sahlane (2012) manualy analysed fallacies in newswire editorials in major U.S. newspapers before invading Iraq in 2003. These two works rely on a list of several fallacy types, such as *ad hominem*, *ad populum*, *appeal to guilt*, *slippery slope*, *hasty generalization*, and few others.

When scaling up annotations and resource acquisitions, serious games provide an alternative to paid crowdsourcing. Recent successful applications include knowledge base extension (Vannella et al., 2014), answering quizes related to medical topics (Ipeirotis and Gabrilovich, 2014), word definition acquisition (Parasca et al., 2016), or word sense labeling (Venhuizen et al., 2013); where the latter one resembles a standard annotation task with bonus rewards rather than a traditional entertaining game. Niculae and Danescu-Niculescu-Mizil (2016) built a game for guessing places given Google Street View images in order to collect data for investigating constructive discussions. An important aspect of serious games for NLP is their benefit to the users other than getting the annotations done quickly: learning a language in *Duolingo*[4] has more added value than killing zombies (despite its obvious fun factor) in *Infection* (Vannella et al., 2014).

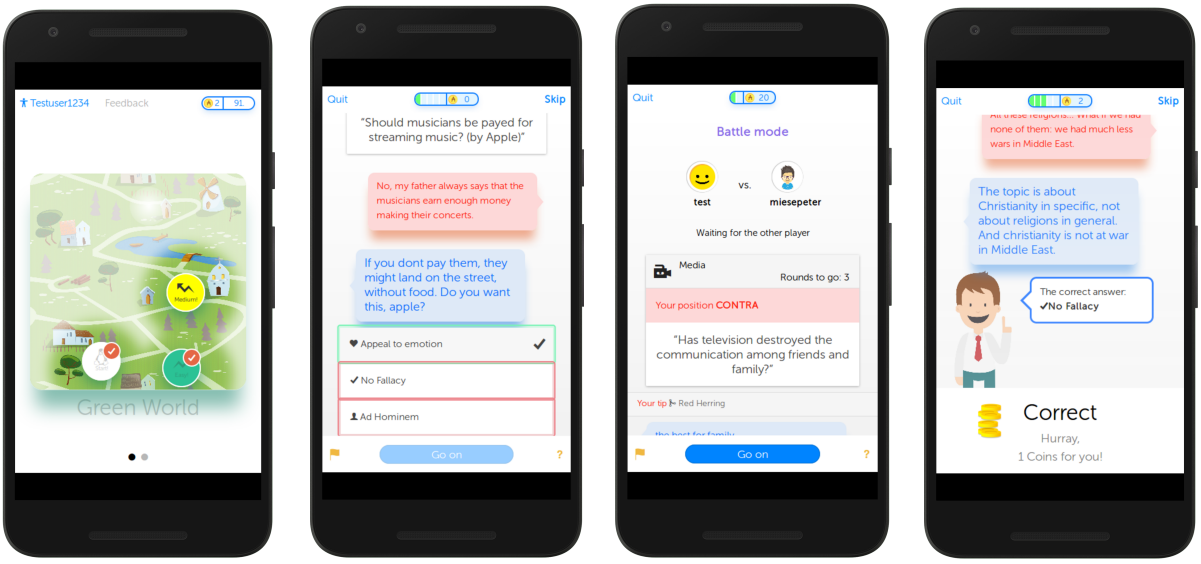
3 Argotario: Overview

Architecture and Implementation *Argotario* is a client-server Web-based application that runs in all modern browsers and seamlessly works on smartphones, providing an authentic look-and-feel. Its three-tier architecture consists of a backend MongoDB database, a Python server behind an Apache2 SSL proxy, and a Javascript client built on top of Ionic framework. *Argotario* is **modular** as it allows developers to add new content (worlds, levels, rounds) as independent modules. The game workflow is **configurable** using JSON files, so it can be customized for evaluating new game scenarios. **Security** is ensured by a SSL certificate and securely hashing all passwords with salt. **Localization** utilizes the built-in capabilities of `ng-translate` so that all texts are stored externally in a JSON file and adding another language to the UI requires only manual translation of these texts.[5] Currently, *Argotario* is available in English

[3] Hamblin (1970) criticized the 'standard treatment' of fallacies widely present in contemporary textbooks as being 'debased,' 'worn-out', 'dogmatic' and 'without a connection to modern logic'.

[4] Although Duolingo presents itself as a learning tool, its incentives and competition features make it feel like accomplishing quests in a game.

[5] Needless to say that providing an initial *content* for a new language, such as a list of language-dependent topics, arguments, and fallacies, requires substantial manual work.



(a) A single *world* with the two first *levels* finished, the third one about to be played, and other to be 'explored'.

(b) The recognize fallacy type *round*.

(c) The *player vs. player* level, now waiting for the opponent's turn.

(d) An example of *hard feedback* in a fallacy recognition round.

Figure 1: Screenshots of *Argotario* taken in a smartphone emulator.

and German.

Game Design We first present the abstract architecture; concrete examples follow in §4. According to Salen and Zimmerman (2004, p. 50), a game is a system consisting of different types of interacting entities that have certain attributes. *Argotario* follows this structure by a hierarchy of **worlds**, **levels**, and game **rounds** (Hannemann, 2015).

A game **round** represents an atomic mini-game in which users take an action and are rewarded with points. Conceptually, each game round follows the same procedure: the users are first faced with *game data*, which they need to interact with. Their response (a choice or free-text input) is then validated with respect to the current game round configuration, similar to form validation on web pages. If the game determines correctness of the response data, it rewards the user with a certain number of points.

A sequence of game rounds form a **level**. To complete a level, all game rounds must be finished, independently of whether the user successfully fulfilled the respective task or not. Whereas game rounds can be re-used in different levels, each level is unique and can be individually designed to fit a certain purpose (i.e., only some types of fallacies are dealt with).

Finally, all levels reside in a **world** which is a wrapper for all included levels, visually resembling a treasure map (see Figure 1a). Their look can be freely customized to be visually appealing and capture a certain atmosphere or theme. There are multiple worlds within the game next to each other.

Users are represented as small circular comic faces (avatars). The first user's **goal** is to finish all levels in all worlds. Initially, the game worlds are covered by a fog, which can be cleared by the user by completing levels. **Ranking** (score) is the second important game goal. Repeating levels allows users to collect more points and hence improve their global rank.

4 Gamifying Fallacy Recognition

The backbone principle of *Argotario* can be summarized as follows. First, since a fallacious argument is one 'that seems to be valid but is not so' (Hamblin, 1970), users must try to 'fool' other users by **writing** a fallacious argument of a given type without being revealed that this is in fact a fallacy. By writing a fallacious argument on purpose with the aim to 'disguise' it as a valid argument, users get sensitive to the very gist of fallacious argumentation (such as rhetorical strategies, linguistic devices, logic, etc.). Second, users learn to **recognize** fallacies in existing arguments—either by revealing the correct fallacy type or stating that the given argument is not fallacious—and get feedback about

their 'debunking' skills (see Figure 1b).[6]

In the serious-game terminology of von Ahn and Dabbish (2008, p. 61), recognizing the correct fallacy type combines the *inversion-problem game* (the guesser produces the input that was originally given to the describer') and a modification of the *output-agreement game* (the guesser has to produce the same output as the crowd; details will be discussed later in §4).

Fallacy Types We gathered an inventory of fallacy types suitable for our game scenarios. Given the breadth and variety of fallacy types (Tindale, 2007; Govier, 2010), we conducted several pilot studies to identify types that are (1) common in everyday argumentative discourse, (2) are distinguishable from each other, and (3) have increasing difficulty. The fallacy type inventory in *Argotario* currently contains *ad hominem*, *appeal to emotion*, *red herring*, *hasty generalization*, *irrelevant authority*, and a non-fallacious argument (Pollak, 2016).

Players learn to recognize different fallacy types gradually, as they accomplish each level. After finishing the first world in which all fallacy types are mastered, users can engage in the *player versus player* world. Here, a dialogue exchange about a given controversy requires users to write fallacious arguments (as in the previous world) and guess which fallacy was used by its opponent (thus getting points for correct answers; details about gold data estimation are explained in the next section). This level is asynchronous; when a user writes a new argument, his opponents gets notification about the turn change, so they do not have to play at the same time (see Figure 1c).

Gold Label Estimation Because all content is created within the game by players with different abilities to write or comprehend argumentation, we treat the data as *noisy* in the first place. First, spam can be reported in all rounds and is submitted to the admins to take action. Second, we rely on MACE (Hovy et al., 2013) for gold label estimation which we seamlessly integrated to the backend. For example, if the user has to write an argument of a given fallacy type, we treat the type only as a single 'vote' and require another four players to guess the correct type of this fallacy in other levels. Only arguments that receive at least five 'votes' are fed into MACE to establish their gold label.

By utilizing crowd voting and spam reporting, we indirectly aim for high-quality labels. Predicting gold labels can be further parametrized by a *threshold* in MACE, which then provides only gold label estimates for instances whose entropy is below the threshold (Hovy et al., 2013, p. 1125). However, a deep analysis of the data quality is on our current research agenda.

Feedback and Incentives *Argotario* provides two types of feedback: *soft* and *hard* one. For labeling arguments with yet unknown gold label, users get only one point without knowing whether their answer was right (*soft feedback*). For arguments with already estimated gold labels, *hard feedback* (see Figure 1d) is given: if the user makes an error, she receives no reward. Apparently, hard feedback is better from the educational point of view as one knows immediately whether her answer was right or wrong; however, users do not know in advance whether a current assessment gives them a soft or hard feedback, so they are inherently encouraged to try their best.

We also built in several sorts of incentives to keep the player engaged. First, *Argotario* shows the overall leaderboard as well as *weekly* ranks to ensure newcomers have chances to succeed, see (Ipeirotis and Gabrilovich, 2014) for details. Players of the week are publicly shown and receive a small monetary prize. Second, debunking fallacious arguments to familiar topics is reportedly entertaining for players interested in rhetoric, argumentation, or public deliberation, according to user feedback obtained after few classroom runs.

5 Benchmarking

So far we tested *Argotario* in serveral user studies and beta-testing sessions. The first study on early versions of *Argotario* examined the effect of hard feedback and the lack thereof on overall users' engagement in the game. We found that the soft feedback has no significant negative impact on the users' experience[7] (Hannemann, 2015).

In a subsequent study, we benchmarked the player vs. player level using Amazon Mechanical Turk (AMT). We asked workers to play a specially configured version of *Argotario* in order to 'win' 20 points required for submitting the HIT. As the player vs. player round needs two dialogue turns of

[6] All written texts and user input are licensed under CC-BY.

[7] Two user groups (20 and 17 participants, respectively) with the same game configuration but with either only soft or hard feedback; final questionnaire with Likert-scale questions; Mann-Whitney-U non-parametric test.

two users and thus two or more people actively participating over a longer period of time, we also implemented a naive *bot* for this study.[8] At the same time, we promoted the game on social media and attracted some non-paid users. Using this process, we could quickly test the entire game mechanism with a larger crowd, identify potential drawbacks, and gather about 1,160 hand-written fallacious arguments. We also experimented with various price per HIT ($1–$2) with respect to average playing time. While the number of rejected low-quality HITs remained negligible for all configurations, we did not observe any correlation between HIT prices and playing times (≈ 18–26 min). Our interpretation is that the HIT price for benchmarking studies should be fair and reflect the study time but does not influence the quality (Pollak, 2016).

6 Conclusions and Outlook

Argotario is a serious game that serves several purposes. First, it is a software tool for computational linguistics research, as it focuses on fallacies in argumentative discourse, an important part of qualitative criteria in computational argumentation. Second, it is software supporting learning and education. Its main educational purpose is to raise awareness—not only that fallacies do exist but they might be easily overlooked and misused in everyday argumentation. Finally, *Argotario* is also a data-acquisition and annotation tool that applies successful techniques for quality estimation from crowd-sourcing approaches. All content is created by users within the game, as opposed to usual annotation tools.

In the long run, we expect that *Argotario* provides a feasible method for data acquisition as compared to standard crowdsourcing. First, a purely monetary-driven perspective is not always the deciding factor of playing additional levels, as shown by Eickhoff et al. (2012). Second, 'experts' from the crowd motivated by the potential for achievement can help engage in participation (Ipeirotis and Gabrilovich, 2014).

[8] We trained a fallacy classifier system on existing arguments in the database using a Convolutional Neural Network based on GloVe embeddings (Pennington et al., 2014) and Keras framework, so the *bot* tried to recognize a fallacy in its opponent arguments during the player vs. player discussion. For generating an answer, it simply looked up an existing fallacy to the given topic. On the one hand, it disobeyed the discourse flow, as it obviously did not coherently respond to its opponent. On the other hand, it allowed us to deploy the game as a HIT on AMT and get a sufficient number of player vs. bot games in a short time.

In the current version, *Argotario* is still a proof of concept. Its capabilities need to be verified at a large scale in order to reveal patterns in the game dynamics with impact on the overall user experience and quality; these cannot be easily anticipated on small-scale benchmarks (§5). In this regard, any manual intervention (such as spam removal) needs to be automated.

Argotario is accessible at `www.argotario.net` along with tutorial videos and runs in any modern web-browser, preferably on smartphones. It is also open-source, source codes under ASL license are available at GitHub.[9]

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), by the GRK 1994/1 AIPHES (DFG), and by the ArguAna Project GU 798/20-1 (DFG).

References

Luis von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Communications of the ACM*, 51(8):57.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The Fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life. *Argumentation*, 29(4):431–456.

Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *ACM SIGIR*, pages 871–880, New York, USA. ACM Press.

Trudy Govier. 2010. *A Practical Study of Argument*, 7th edition. Wadsworth, Cengage Learning.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *EMNLP*, pages 1214–1223, Austin, Texas.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *ACL (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

[9] `https://github.com/UKPLab/argotario`

Charles L. Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Raffael Hannemann. 2015. Serious games for large-scale argumentation mining. Master Thesis, Techische Universität Darmstadt.

Hans Hansen. 2015. Fallacies. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2015 edition. Metaphysics Research Lab, Stanford University.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *NAACL-HLT*, pages 1120–1130, Atlanta, Georgia.

Panagiotis G. Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: Targeted Crowdsourcing with a Billion (Potential) Users Panagiotis. In *WWW'14*, pages 143–154, Seoul, South Korea. ACM.

David Jurgens and Roberto Navigli. 2014. It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL*, 2(1):449–463.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, pages 787–798, Doha, Qatar.

Igor Mayer, Geertje Bekebrede, Casper Harteveld, Harald Warmelink, Qiqi Zhou, Theo van Ruijven, Julia Lo, Rens Kortmann, and Ivo Wenzler. 2014. The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, 45(3):502–527.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Zuzana Nevěřilová. 2014. Annotation game for textual entailment evaluation. In *CICLing*, pages 340–350, Kathmandu, Nepal.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *NAACL-HLT*, pages 568–578, San Diego, CA, USA.

Petteri Nieminen and Anne-Mari Mustonen. 2014. Argumentation and fallacies in creationist writings against evolutionary theory. *Evolution: Education and Outreach*, 7(1):11.

Fabio Paglieri. 2016. Don't worry, be gappy! On the unproblematic gappiness of alleged fallacies. In *The psychology of argument*, chapter 9, pages 1–20. College Publications, London.

Iuliana-Eena Parasca, Andreas Lukas Rauter, Jack Roper, Aleksandar Rusinov, Guillaume Bouchard, Sebastian Riedel, and Pontus Stenetorp. 2016. Defining Words with Words: Beyond the Distributional Hypothesis. In *RepEval*, pages 122–126, Berlin, Germany.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, Doha, Qatar.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. on Interactive Intelligent Systems*, 3(1):1–44.

Christian Pollak. 2016. Serious games for learning fallacy recognition. Master Thesis, Techische Universität Darmstadt.

Ahmed Sahlane. 2012. Argumentation and fallacy in the justification of the 2003 War on Iraq. *Argumentation*, 26(4):459–488.

Katie Salen and Eric Zimmerman. 2004. *Rules of Play: Game Design Fundamentals*. MIT Press.

Edward Schiappa and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*, 1st edition. Pearson UK.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56, Doha, Qatar.

Christian Stab and Iryna Gurevych. 2017. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *EACL (Volume 1, Long Papers)*, pages 980–990.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*, critical reasoning and argumentation edition. Cambridge University Press, New York, NY, USA.

Frans H. Van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 2014. Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *ACL (Volume 1: Long Papers)*, pages 1294–1304, Baltimore, Maryland USA.

Noortje J. Venhuizen, Valerio Basile, Kilian Evang, and Johan Bos. 2013. Gamification for Word Sense Labeling. In *IWCS (Short Papers)*, pages 397–403, Potsdam, Germany.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, page (to appear). Association for Computational Linguistics.

Douglas Walton. 1995. *A Pragmatic Theory of Fallacy*. The University of Alabama Press, Tuscaloosa, AL.