# Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices

**Ivan Habernal**[†‡], **Patrick Pauli**[†], **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab
[‡]Research Training Group AIPHES
Department of Computer Science, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany
www.ukp.tu-darmstadt.de www.aiphes.tu-darmstadt.de

**Abstract**

As argumentation about controversies is culture- and language-dependent, porting a serious game that deals with daily argumentation to another language requires substantial adaptation. This article presents a study of deploying *Argotario* (serious game for learning argumentation fallacies) in the German context. We examine all steps that are necessary to end up with a successful serious game platform, such as topic selection, initial data creation, or effective campaigns. Moreover, we analyze users' behavior and in-game created data in order to assess the dissemination strategies and qualitative aspects of the resulting corpus. We also report on classification experiments based on neural networks and feature-based models.

**Keywords:** argumentation, fallacies, serious games

## 1. Introduction

Computational argumentation and argument mining has been traditionally dealing with understanding argument's structure (Habernal and Gurevych, 2017; Stab and Gurevych, 2017; Eger et al., 2017; Daxenberger et al., 2017). Recently, attention has been paid to pragmatic aspects of arguments, such as convincingness (Habernal and Gurevych, 2016b; Habernal and Gurevych, 2016a) or overall quality (Wachsmuth et al., 2017a). A fairly unexplored area of computational argumentation is *fallacies*: arguments that seem to be valid but are not so (Hamblin, 1970). To tackle the nonexistence of corpora for dealing with fallacies computationally, Habernal et al. (2017) published *Argotario*—a serious game intended to educate players and at the same time collect annotated fallacious arguments.

The majority of research on computational argumentation is English-centric (with several exceptions, such as (Peldszus and Stede, 2015; Liebeck et al., 2016; Chow, 2016)). Not only the language itself but the discussed topics and controversies are culture-specific. For example, 'homeschooling' or 'death penalty' are almost non-existent in Germany, while being highly controversial subjects of discussion in the United States. As *Argotario* had been developed within the English context, simply translating the topics and existing arguments and fallacies into another language does not meet the expectations of a serious game user.

We thus asked the following research questions. First, what are the best means to tackle language adaptation in serious games that depend on world-knowledge, cultural context, and specific controversies (RQ1)? Second, we were interested in the dynamics and outcomes of deployment of the game, namely whether new lay users understand differences between various fallacies (RQ2), which qualitative aspects are to be expected in user-written fallacies (RQ3), and which advertising channels deliver the best return of investment (RQ4).

To answer these questions, we added German language support to *Argotario* and crowd-sourced initial data to face the 'cold-start' problem (RQ1), launched several campaigns (RQ4) and analyzed the obtained data and users' behavior (RQ2; RQ3). Moreover, we conducted several experiments with feature-based and deep-learning models for classifying fallacies. The main contributions of this article are (1) an extensive study of language adaptation of a fallacy-oriented serious game and (2) a dataset released to the community under CC-0 license which is, to the best of our knowledge the first corpus of German and English fallacious arguments.

## 2. Related work

Fallacies have been thoroughly studied in argumentation theory (Damer, 2013; Tindale, 2007; Schiappa and Nordin, 2013; Walton, 1995; van Eemeren et al., 2014). Despite the vast number of theoretical approaches, empirical research and analysis of fallacies in actual argumentative discourse has been rather limited in scope and size. Several recent endeavors in that direction include, e.g., a manual examination of fallacies found in articles supporting creationism by (Nieminen and Mustonen, 2014) or a manual analysis of fallacies in newswire editorials in major U.S. newspapers before invading Iraq in 2003 by (Sahlane, 2012). These examples demonstrate the enormous persuasive effect of fallacious argumentation; other examples of its rhetorical power can be found in (Macagno, 2013).

The computational perspective on fallacies in natural language arguments has been bound to the process of obtaining reliable data from the crowd and serious-game players (Pollak, 2016; Habernal et al., 2017). There are also several related works devoted to argumentation quality, such as confirmation bias (Stab and Gurevych, 2016), or qualitative assessment of arguments from the Web (Wachsmuth et al., 2017b).

## 3. Overview of Argotario

*Argotario* represents an instance of so-called serious games (Mayer et al., 2014) that deals with fallacies in everyday argumentation. *Argotario* is an open-source, platform-independent application with strong educational aspects.[1] It

[1] www.argotario.net

was primarily developed in English but has been extended to support multiple languages (Habernal et al., 2017). In short, players of *Argotario* learn to recognize several types of fallacies as well as to write them, both in a single-player and player vs. player scenarios (an example of an actual player vs. player round is shown later in Figure 2). All arguments[2] are thus composed and evaluated in-game with a minimal intervention, except of the initial data and topic selection.

### 3.1. Fallacy inventory

Labeling an argument as a fallacy of a certain type is usually clear in textbook examples only (Tindale, 2007; Govier, 2010). While several taxonomies exist in the literature, their empirical usefulness is usually not warranted (Boudry et al., 2015). We thus approached the selection of fallacy types using a bottom-up approach such that the inventory contains fallacy types that are distinguishable from each other and are common in everyday discourse. After several pilot studies, the final inventory consists of the following fallacy types; the examples are actual fallacies written by *Argotario* players.

**Ad hominem** The opponent attacks a person instead of arguing against the claims that the person has put forward. *Example:* "Yeah, and you are a guy who loves war, that's it. You like it when people die." (Topic: Should the fight versus the Islamic State include military operations?

**Appeal to emotion** This fallacy tries to arouse non-rational sentiments within the intended audience in order to persuade. *Example:* "Yes, all the polar-bears are dying, and we are next." (Topic: Is global warming really an issue?)

**Red herring** This argument distracts attention away from the thesis which is supposed to be discussed. *Example:* "I am a hunter. Animals need to die in order to keep balance in the forest." (Topic: Should we allow animal testing for medical purposes?)

**Hasty generalization** The argument uses a sample which is too small, or follows falsely from a sub-part to a composite or the other way round. *Example:* "Yes, Facebook is censoring racist comments against refugees. It works quite well. All media should be censored." (Topic: Is it effective to censor parts of the media?)

**Irrelevant authority** While the use of authorities in argumentative discourse is not fallacious inherently, appealing to authority can be fallacious if the authority is irrelevant to the discussed subject. *Example:* "Yes, my husband has the same opinion." (Topic: Is television an effective tool in building the minds of children?)

**Non-fallacious argument** None of the above. Note that we don't use the term "valid" or "good" argument due to the inherent subjective evaluative meaning of these adjectives.

[2]We will use *arguments* and *fallacies* to refer to the same concept here, namely a (potentially fallacious) argument.

### 3.2. Game design

In particular, the game is structured into *game rounds* (an atomic mini-game in which an interaction from the user is required and is usually rewarded with points), *levels* (predefined sequences of game rounds, for example with increasing difficulty), and *worlds* (a set of interconnected levels that are visualized as a landscape in the game).

**Example** One concrete example of a game round is the following: The user is shown an argument to a given topic and her quest is to guess whether the argument is fallacious and if so, which fallacy was committed (fallacy type classification, in other words). Let's assume the correct answer is known to the system.[3] The user is then awarded a point if answered correctly, or no reward is given otherwise. Other game rounds include, for instance, writing a fallacious argument given the topic and the intended fallacy type.

The pre-defined levels are either educative, thus gradually teaching the user all fallacy types using the above-mentioned fallacy classification rounds and writing rounds, or they require two players competing against each other.
The users have to achieve two goals. First, they must finish all predefined levels in the first worlds and therefore learn all fallacy types. The second goal is to achieve high ranking (overall score). The players' scores are shown on a global leaderboard with weekly and overall scores. Scoring high in the weekly leaderboard is another incentive to motivate the player. To get a better sense of the gameplay, we recommend watching the videos at `www.argotario.net`.

## 4. Porting Argotario to German

As *Argotario* comes with a set of predefined controversial topics, the content has to be interesting and relevant enough to grab players' attention from the beginning and at the same time diverse enough to remain entertaining for long-term players. Furthermore, the target group has to be taken into account. Therefore, translating the English topics into German results into mismatch due to different culture-related controversies.
We thus obeyed the following criteria when selecting topics for the German version, namely (a) presence in German mass media, (b) relevance for a political point of view as articles on politics are generally among the most commented ones in online newspapers and the comment sections are full of fallacious arguments, and (c) long-term orientation which would filter out short-term political scandals or quickly abating trends. We manually compiled a list of 30 topics that fit the criteria; see Appendix A for the full listing.
Another 'cold-start problem' is the need of an initial set of fallacious arguments. In *Argotario*, players have to first learn to recognize existing fallacies, before they are asked to write new ones. Therefore the initial set of arguments is important for the first impression of the game to a new user. We opted for paid crowdsourcing on the German platform Clickworker to collect fallacious arguments for all the 30 argumentation topics. Workers were paid €1.50 for writing three arguments that are pro or con to the given topics. As

[3]The 'correct answers' (gold labels) are estimated based on users' voting, see (Habernal et al., 2017) for details.

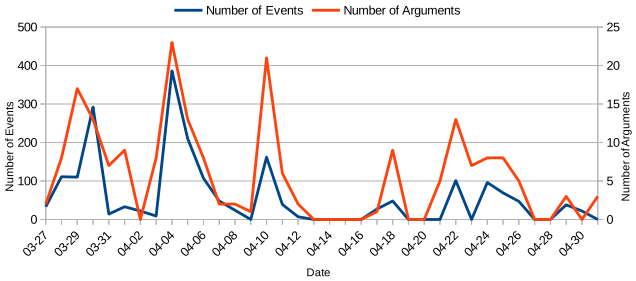


Figure 1: Game activity over time during the evaluation period

a result, we collected 90 high-quality[4] German arguments, which is sufficient for launching the game publicly.

We compared the price with English crowdsourced fallacies by Pollak (2016) who used Amazon Mechanical Turk (AMT). In his experiments, an average price was $0.15 per argument. However, our experiments with AMT were unsuccessful due the German-fluency requirements on the workers. It is thus significantly more difficult and expensive to crowdsource arguments for the game in languages other than English.

## 5. Data collection campaigns

During the evaluation period between March and April 2017, the game was advertised using three different channels: (1) personal recommendations by the authors, (2) postings advertising *Argotario* in internet forums about politics, language learning, and philosophy, and (3) a one-week paid Google AdWords campaign. These channels were used sequentially which allowed us to measure the success of each one individually. Overall, the majority of players (69%) played the game just once after registration and did not return later. But there is also a group of players who returned several days after the registration.

### 5.1. Players' activity

We logged several *events*, such as that a player entered a world, finished a world, started playing a level, or finished a game round. Figure 1 shows the number of logged events (in blue) and the number of written arguments (in red) over time during the five weeks of the evaluation period. The three phases distinguished previously are clearly visible in both graphs. The first peak (until April 2) is produced by players who followed personal recommendation (P1), the second and third peaks by players attracted by forum postings (P2, until April 13). The activity during the last two weeks is caused by the AdWords campaign (P3).

To further break down the activity of the three player groups (P1–P3), Table 1 summarizes the average and the distribution of the playing time span as well as the average number of arguments and judges. The player group P2 contains the most long-term users. 12% played the game for more than one day and 15% for at least two days with an interval of more than five days, which makes a percentage of 27% who were sufficiently attracted by the game to return to playing it. This observation is also affirmed by the average number of arguments written by one player which is shown in the penultimate column. As a result, it is clearly visible that the players attracted by forum postings have been the most active out of the three player groups.

### 5.2. Google AdWord campaign

Advertising using Google AdWords has already been applied in NLP research as a technique to attract users for annotating applications and games (Ipeirotis and Gabrilovich, 2014).

We launched the AdWords campaign for *Argotario* on April 20, 2017 with a €50 budget. Table 2 shows the performance of keywords in terms of clicks on the advertisement, as well as the number of impressions with this keyword (how many times the advertisement was displayed), and the Click-Through-Rate (CTR), which is the probability that a user clicks on the advertisement. Overall, the CTR is rather low as compared the total number of impressions.

All in all, the Google AdWords attracted only 9 players to register and play Argotario (from the 145 clicks), which is rather unsatisfying; the costs are about €5.5 per registered user. The effectiveness of advertisements could be improved by better analytics and targeting (such as in (Ipeirotis and Gabrilovich, 2014), who relied on the Google Analytics tool), or a more fine-grained keyword selection. The campaigns should rather advertise with a high investment for a shorter period than with low investment spread over a longer time span.

## 6. Data analysis

During the evaluation phase we collected 296 German arguments which we analyzed from different perspectives.

### 6.1. Language properties and quality

The arguments are on average 18.2 words long, and most of them consist of one or two sentences. This corresponds to findings of Best (2002) who showed that in German journalistic texts the sentence length ranges between 9.62 words (average length of the shortest sentence) and 22.91 words (average length of the longest sentence). Also, Pieper (1979) states that the median sentence length in German discussions is 11.83 words. This comparison shows that the collected arguments share the characteristics of German discussion texts.

Another aspect that serves as an indicator of how serious the game is taken by the players is the quality of orthography and grammar. We manually analyzed spelling mistakes (counted as orthographical errors) as well as grammatical errors (such as wrong case endings of nouns and adjectives, wrong usage of capitalization, missing punctuation or words, and wrong word order). Most arguments (85%) were completely error free, only a fraction contained orthographical errors (7%) or grammar errors (11%).

### 6.2. Discourse properties

One of the goals of *Argotario* is to collect a corpus of fallacies that resembles a typical Web discussion. This refers particularly to the arguments which are part of the Player

[4] In terms of whether they fit the requested fallacy type and whether they are 'fallacious' enough; a manual analysis was done by the authors.

| | | | Playing Time Span (PTS) | | | | Activity Level | |
|---|---|---|---|---|---|---|---|---|
| | #Players | Avg. PTS | <1h | <1d | <5d | >5d | Avg. #Args | Avg. #Judges |
| P1 | 23 | 1d 7h 46min | 74% | 17% | 4% | 4% | 3.74 | 7.48 |
| P2 | 26 | **2d 11h 58min** | 69% | 4% | 12% | **15%** | **4.65** | **13.81** |
| P3 | 9 | 0d 4h 14min | 56% | 33% | 11% | 0% | 3.56 | 6.56 |
| Total | 58 | 1d 16h 8min | 69% | 14% | 9% | 9% | 4.12 | 10.17 |

Table 1: Playing Time Span and average number of arguments and judges per player group

| Keyword | Clicks | Impressions | CTR (%) |
|---|---|---|---|
| **Total** | 145 | 31871 | 0.45 |
| argumentieren (*argue/arguing*) | 24 | 733 | 3.27 |
| deutsch argumentieren (*arguing in German* ) | 24 | 342 | 7.02 |
| politik (*politics*) | 19 | 9897 | 0.19 |
| lernspiel (*educational game*) | 14 | 5659 | 0.25 |
| bundestagswahl (*general election in Germany*) | 11 | 1374 | 0.80 |
| lernspiel online (*online educational game*) | 10 | 2513 | 0.40 |
| online lernspiel (*online educational game*) | 7 | 2670 | 0.26 |
| philosophie (*philosophy*) | 5 | 4869 | 0.10 |
| argumentieren lernen (*learn how to argue*) | 3 | 53 | 5.66 |
| duolingo | 2 | 349 | 0.57 |

Table 2: Keywords statistics of the Google AdWord campaign for Argotario; top 10 'clicks' keywords are shown

**A (Hasty Generalization)**: *"Da ohnehin alle Sportler dopen, hat ein Verbot keine Auswirkungen."* (As all athletes dope anyway, the prohibition has no effects.)
**B (Ad Hominem)**: *"Sie sind doch ein Zyniker, wenn Sie das behaupten."* (You are a cynic when you assert this.)
**A (Appeal to Emotion)**: *"Das hat mit Zynismus nichts zu tun - Solange wir so tun, als hätten wir ein funktionierendes System, solange leiden ehrliche Sportler massiv unter dieser Ungerechtigkeit. Das muss man sich immer wieder vor Augen führen, wer hier leidtragend ist!"* (This has nothing to do with cynicism - As long as we act as if we had a functional system, honest athletes suffer massively under this unfairness. You always have to visualize who is the bereaved here!)
**B (Red Herring)**: *"Und deswegen soll man die Dopingkontrollen abschaffen? Wollen Sie dann auch alle Steuerprüfungen abschaffen, weil es momentan kein perfektes System gibt, mit dem man alle Steuersünder erwischt?"* (And that is why doping controls should be abolished? Do you then also want to abolish tax inspections because there is no perfect system at the moment with which all tax evaders are caught?)

Figure 2: Example dialog about doping in sports in a Player vs. Player round (A vs. B); parentheses show the desired fallacy type the player was instructed to compose.

vs. Player (PvP) round (42.2% of the corpus under investigation). One important feature is discourse coherence, in particular the presence of arguments that directly respond to the opponent's last argument. We define an argument as *dialogical* if the player refers to the opponent's argument and *monological* if the player just writes a stand-alone context-independent argument. Manual analysis revealed that 63% of arguments in the PvP round are dialogical. This is a satisfying result because players are not explicitly instructed to obey any discourse coherence and are only asked to compose a particular type of fallacious arguments. Figure 2 shows one PvP game about doping in sports with clear dialogical properties.

## 6.3. Fallacy type accuracy

The educational objective of *Argotario* is to teach players what a fallacy is and which fallacy types exist, in particular, players have to compose arguments of the given fallacy type. Analyzing the arguments written by the players contributes to answering the question to what extent this objective is reached. We investigated how accurate the written fallacies are by manually re-labeling the full set of 296 arguments (which we will call *expert fallacy type*). The originally requested fallacy type to be composed by the author is referred to as *intended fallacy type*. Moreover, players also 'judge' other players' fallacies. When at least three labels (votes) for the same argument are available, a gold standard label is estimated (which we call *voting fallacy type*); see (Habernal et al., 2017, p. 10) for details. These labels are available for a subset of 92 arguments that received three and more votes during our evaluation period.

**Intended fallacy vs. expert fallacy** The intended fallacy type corresponds with the expert fallacy type on 229 of the 296 arguments (average macro F1 score 77%). The results vary widely depending on the fallacy type. While the F1 score for *irrelevant authority* reaches 95%, *red herring* results are around 60%. We can conclude from this that, at least in the context of the game, composing arguments which distract the attention is more difficult in contrast to arguments that use irrelevant authority as backing; *red herring* arguments demand more creativity as it is not as trivially possible to refer to the previous contextual statements.

**Voting vs. expert fallacy** We further examined whether the 'collective intelligence' through in-game voting leads to accurate labels by comparing the result with the expert fallacy types. As many as 90 of 92 judged arguments (97.8% accuracy) match the correct expert fallacy label. We investigated the two wrongly predicted arguments and found that they were ironical response to the previous argument which features the fallacy of *irrelevant authority* and contains a

| Fallacy type | Instances | (in %) |
|---|---|---|
| No fallacy | 123 | 28 |
| Appeal to emotion | 101 | 23 |
| Red herring | 52 | 12 |
| Ad hominem | 62 | 14 |
| Hasty generalization | 38 | 9 |
| Irrelevant authority | 54 | 13 |

Table 3: Class distribution for classification experiments

| Model | Accuracy (%) | Macro-F1 (%) |
|---|---|---|
| Random-guess baseline | 19.6 | 16.7 |
| Majority class baseline | 28.6 | 7.4 |
| Bi-LSTM | 50.9 | 42.1 |
| SVM | 46.3 | 37.2 |

Table 4: Overview of the classification results

reference to German history. Three players have judged the arguments and were completely in disagreement. It is debatable if a (non-explicit) reference to Germany's war history of the 20th century is to be qualified as *appeal to emotion* (because it maybe intends to evoke guilt and shame) or as a valid argument. This is a good example for an argument where the correct fallacy type is not 100% clear and cannot be reliably decided by a single annotator. Overall, we can conclude that the wisdom of the crowd works here with very high reliability and even with only three votes it is possible to estimate the true gold label, correct the author of the original argument, and maintain a high quality. This is important for giving the right feedback to novice users when recognizing fallacies.

## 7. Classification experiments

As one of the long-term goals of *Argotario* is to provide training data for automatic fallacy recognition, we were interested to which extent this problem is solvable using the data gathered so far. The chosen NLP methods are not new but they do reflect the mainstream approaches to classification using deep learning (Goldberg, 2017), thus provide meaningful baselines for further endeavors.

This section thus sketches some classification experiments we performed on the German fallacy dataset. We use all arguments collected during the evaluation period, the arguments created as start-up data, and arguments collected during pilot testing. All arguments were re-labeled by the authors (similarly to *expert fallacy type*) to ensure their reliability. This resulted into 430 labeled arguments with six classes (*no fallacy*, *appeal to emotion*, *red herring*, *ad hominem*, *hasty generalization*, and *irrelevant authority*).

We experimented with a bi-directional LSTM model (Hochreiter and Schmidhuber, 1997) fed with German 64-dimensional word embeddings (Al-Rfou et al., 2013). As a second model, we opted for Support Vector Machines with a range of manually compiled features (question marks, discourse markers signaling nesting, punctuation, length, capitalization, pronouns signaling attacks, reported-speech words, and few others). All experiments were conducted using 10-fold cross validation. The dataset contains 430 gold-labeled arguments, distributed as shown in Table 3.

Overall results are shown in Table 4. A detailed examination of Bi-LSTM results revealed that the best performance was achieved for *no fallacy* and *ad hominem* (F1 score over 60%). Classes with the lowest F1 scores were *red herring* and *hasty generalization*, partly due to their limited presence in the dataset (12% and 8%, respectively). While SVM performed only slightly worse than Bi-LSTM, the *red herring* class was never predicted; this can be explained by the fact that the SVM features concentrate on formal and lexical cues, wile *red herring* arguments rarely contain signal words or other typical signatures. To better identify *red herrings*, it would be necessary to implement features on higher levels of the NLP pre-processing chain, including semantics and world knowledge.

We further compared our results to experiments with the same classification task on English dataset containing 1,160 arguments (Pollak, 2016). While Pollak (2016) achieved 40% macro F1 using Convolutional neural network (CNN), our CNN model performed worse (37%; not reported in Table 4). On the other hand, we achieved better F1 score using a simple LSTM (47%; not reported in Table 4) than Pollak (2016) (43%). However, these results must be taken with a grain of salt as both datasets are rather small to fully leverage the power of deep neural networks.

## 8. Conclusion and Outlook

We showed that porting a serious game dealing with fallacious argumentation to another language requires substantial effort in adapting topics and preparing high quality starting data (RQ1). We approached the topic selection empirically by relying on mass media and paid crowd-sourcing, which delivered a reasonable starting setup. We manually analyzed the in-game produced data and found that while users are only partially correct in writing fallacies of the given type, the 'wisdom of the crowd' through in-game voting caters for precise corrections and thus high-quality labels (RQ2; Section 6.3.). Furthermore, most users intuitively obey discourse coherence and clear writing (RQ3; Sections 6.1. and 6.2.). By exploring the dynamics of campaigns and user behavior, we found that a high conversion rate of most game rounds shows that players find *Argotario* attractive and do not quit the game early. Out of the three advertising methods, postings in forums have shown the most success (RQ4; Section 5.).

All data from *Argotario* are published under permissive Creative Commons Zero (CC0) license and can be obtained at `https://github.com/UKPLab/argotario`.

## 9. Acknowledgments

This work has been supported by the ArguAna Project GU 798/20-1 (DFG), and by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1).

## 10. Bibliographical References

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192,

Sofia, Bulgaria, August. Association for Computational Linguistics.

Best, K.-H. (2002). Satzlängen im deutschen: Verteilungen, mittelwerte, sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft*, 7:7–31.

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

Chow, M. (2016). Argument Identification in Chinese Editorials. In *Proceedings of the NAACL Student Research Workshop*, pages 16–21, San Diego, CA, USA. Association for Computational Linguistics.

Damer, T. E. (2013). *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Cengage Learning, Boston, MA, 7th edition.

Daxenberger, J., Eger, S., Habernal, I., Stab, C., and Gurevych, I. (2017). What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2045–2056, Copenhagen, Denmark. Association for Computational Linguistics.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*, volume 37 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

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

Habernal, I. and Gurevych, I. (2016a). What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016b). 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)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

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

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational Argumentation Meets Serious Games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Hamblin, C. L. (1970). *Fallacies*. Methuen, London, UK.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ipeirotis, P. G. and Gabrilovich, E. (2014). Quizz: Targeted Crowdsourcing with a Billion (Potential) Users Panagiotis. In *WWW '14: Proceedings of the 23rd international conference on World wide web*, pages 143–154, Seoul, South Korea. ACM.

Liebeck, M., Esau, K., and Conrad, S. (2016). What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld. In *Proceedings of the Third Workshop on Argument Mining*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Macagno, F. (2013). Strategies of character attack. *Argumentation*, 27(4):369–401.

Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., van Ruijven, T., Lo, J., Kortmann, R., and Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, 45(3):502–527.

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

Peldszus, A. and Stede, M. (2015). Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Pieper, U. (1979). *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Gunter Narr, Tübingen.

Pollak, C. (2016). Serious games for learning fallacy recognition. Master Thesis, Techische Universität Darmstadt.

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

Schiappa, E. and Nordin, J. P. (2013). *Argumentation: Keeping Faith with Reason*. Pearson UK, 1st edition.

Stab, C. and Gurevych, I. (2016). Recognizing the Absence of Opposing Arguments in Persuasive Essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118, Berlin, Germany. Association for Computational Linguistics.

Stab, C. and Gurevych, I. (2017). Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

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

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Springer, Berlin/Heidelberg.

Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., and Stein, B. (2017a). Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255. Association for Computational Linguistics.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017b). Com-

putational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 17)*, page (to appear), April.

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

## A List of German argumentation topics

- Es ist gut, dass Deutschland viele Flüchtlinge aufgenommen hat. *Germany accepting that many refugees was good.*
- Geschäfte sollten jeden Tag rund um die Uhr geöffnet sein. *Shops should be open 24-7.*
- Öffentliche Verkehrsmittel sollten kostenlos sein. *Public transportation should be free of charge.*
- Deutschland sollte ein Bedingungsloses Grundeinkommen einführen. *Germany should implement basic income.*
- In der Schule sollte es keine Noten geben. *Schools should abandon grades.*
- Der Konsum von Cannabis sollte legalisiert werden. *Marijuana consumption should be legalized.*
- An Universitäten sollte keine Forschung zu militärischen Zwecken durchgeführt werden. *Universities should not carry out any military research.*
- Man sollte schon ab 16 Jahren wählen dürfen. *Voting age should be reduced to 16.*
- Die Türkei sollte in die EU aufgenommen werden. *Turkey should not join the European Union.*
- Gentechnik ist etwas Gutes. *Genetic engineering is good.*
- Es sollte mehr Videoüberwachung im öffentlichen Raum geben. *There should be more surveillance cameras in public areas.*
- Die Erbschaftssteuer sollte erhöht werden. *The inheritance tax should be raised.*
- Es ist richtig, dass Hartz-IV-Empfängern ihre Leistungen gekürzt werden, wenn sie ein Job-Angebot ablehnen. *Reducing Hartz-IV*[5] *benefits of those who refuse a job offer is right.*
- Homosexuelle Paare sollten Kinder adoptieren dürfen. *Homosexual couples should be allowed to adopt children.*
- Man lebt besser, wenn man vegan lebt. *Vegan lifestyle is a better lifestyle.*
- Der Verfassungsschutz sollte abgeschafft werden. *The Verfassungschutz*[6] *should be abandoned.*
- Der Euro sollte abgeschafft werden. *The Euro currency should be abandoned.*
- Es sollte an staatlichen Schulen keinen Religionsunterricht geben. *Religion classes should not be taught at public schools.*
- Ausländer sollten an Kommunalwahlen teilnehmen dürfen. *Foreigner should be allowed to participate in local elections.*
- Die Wehrpflicht sollte wieder eingeführt werden. *Compulsory military service should be enforced again.*
- Es wird zu wenig gegen den Klimawandel unternommen. *Climate change measures are unsatisfying.*
- Die Agenda 2010 war gut für Deutschland. *The "Agenda 2010"*[7] *was good for Germany.*
- Es sollte eine Steuer auf Plastikverpackungen eingeführt werden. *Plastic packaging should be taxed.*
- Kriminelle Ausländer sollten sofort abgeschoben werden. *Criminal foreigners should be deported right away.*
- Doping im Sport sollte legalisiert werden. *Doping in sport should be legalized.*
- Deutsche Städte sollten sich für die Ausrichtung von Olympischen Spielen bewerben. *German cities should apply for Olympic games.*
- Es ist besser, Bücher zu lesen statt Filme zu schauen. *Reading books is better than watching movies.*
- Es ist gut, dass klassische Musik mit Steuergeld gefördert wird. *Supporting classical music with tax money is good.*
- Man sollte militärisch gegen den IS vorgehen. *Military operations should be taken against the IS.*
- Kinder sollten früh mit Computern und Smartphones umgehen lernen. *Children should learn how to operate computers and smartphones in the early age.*

[5] https://en.wikipedia.org/wiki/Hartz_concept

[6] https://de.wikipedia.org/wiki/Verfassungsschutz

[7] https://en.wikipedia.org/wiki/Agenda_2010