Computational Models of Argument
S. Modgil et al. (Eds.)

doi:10.3233/978-1-61499-906-5-413

Empirical Cognitive Study on Abstract Argumentation Semantics

Marcos CRAMER, Mathieu GUILLAUME
University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract. In abstract argumentation theory, multiple argumentation semantics have been proposed that allow to select sets of jointly acceptable arguments from a given set of arguments based on the attack relation between arguments. The existence of multiple argumentation semantics raises the question which of these semantics predicts best how humans evaluate arguments, possibly depending on the thematic context of the arguments. In this study we report on an empirical cognitive study in which we tested how humans evaluate sets of arguments depending on the abstract structure of the attack relation between them. Two pilot studies were performed to validate the intended link between argumentation frameworks and sets of natural language arguments. The main experiment involved a group deliberation phase and made use of three different thematic contexts of the argument sets involved. The data strongly suggest that independently of the thematic contexts that we have considered, strong acceptance and strong rejection according to the CF2 and preferred semantics are a better predictor for human argument acceptance than the grounded semantics (which is identical to strong acceptance/rejection with respect to complete semantics). Furthermore, the data suggest that CF2 semantics predicts human argument acceptance better than preferred semantics, but the data for this comparison is limited to a single thematic context.

Keywords. abstract argumentation, argumentation semantics, empirical cognitive study

Introduction

In the context of abstract argumentation theory as introduced by [11], multiple *argumentation semantics* have been proposed in the literature as criteria for selecting acceptable arguments based on the structure of the attack relation between the arguments (see [4]). While most of the proposed semantics agree in some simple cases, e.g. in the case of the *simple reinstatement* argumentation framework depicted in Figure 1 on page 4, the different semantics disagree on the acceptability of arguments in more complex cases.

Given that the applicability of abstract argumentation theory to human reasoning is desirable, this situation gives rise to the following three research questions: First, in those cases in which all the standard argumentation semantics agree, do humans actually evaluate the acceptability of arguments as predicted

by these argumentation semantics? Second, in those cases in which there is disagreement between the standard argumentation semantics, which one of them best predicts the judgments that humans make about the arguments? Third, does the answer to these questions depend on the thematic context of the arguments, or is it largely context-independent?

Previous studies on cognitive aspects of abstract argumentation theory have made some limited progress on these research questions, especially the study by Rahwan *et al.* [19]. In this paper, we present the results of an empirical cognitive study which was designed to overcome some of the limitations of their study so as to make progress on all three research questions described above. The experiment involved three different thematic contexts of the argument sets involved. One important feature of our study is that we did not take for granted any assumptions about the directionality of attacks between natural language arguments, but performed two pilot studies to test the perceived directionality of attacks. Another important feature of our study is that it involved a group deliberation phase, which has been shown to increase performance in logical reasoning tasks [13,2].

Many standard semantics allow for multiple conflicting extensions or labellings, which represent different internally consistent judgments about the arguments. In our experiment, however, participants were asked to make a single judgment about each argument, so we compare their judgments to the *justification status* of arguments according to various semantics (see [24,4]), as the justification status is always unique. In particular, we focus on the justification statuses *strong accept* and *strong reject*, which in the labeling approach to argumentation semantics correspond to always being labeled `in` or always being labeled `out`, respectively. The justification status in complete semantics directly corresponds to grounded semantics, so we do not consider complete semantics separately.

The data from our experiment strongly suggest that independently of the thematic contexts that we have considered, strong acceptance and strong rejection according to the CF2 and preferred semantics are a better predictor for human argument acceptance than the grounded semantics. Furthermore, the data suggest that CF2 semantics predicts human argument acceptance better than preferred semantics, but the data for this comparison is limited to a single thematic context.

1. Preliminaries of Abstract Argumentation Theory

We will assume that the reader is familiar with the basics of abstract argumentation theory as introduced by Dung [11] and as explained in its current state-of-the-art form by Baroni et al. [4]. In particular, we will assume that the reader knows the notion of an *argumentation framework* (AF) as well as the *complete*, *grounded*, *stable* and *preferred* argumentation semantics as introduced by Dung [11], both in their traditional extension-based variant and in their label-based variant [3,4]. In this section we furthermore define CF2 semantics as well as the notions of *strong acceptance* and *strong rejection*, and we show how all these notions can be applied to the three AFs that we will refer to in later sections. For reasons that are explained at the end of this section, in this paper we focus on grounded, preferred and CF2 semantics.

CF2 semantics was first introduced by Baroni *et al.* [5]. The idea behind it is that we partition the AF into *strongly connected components* and recursively evaluate it component by component by choosing maximal conflict-free sets in each component and removing arguments attacked by chosen arguments. We formally define it following the notation of Dvořák and Gaggl [12]. For this we first need some auxiliary notions:

Definition 1. Let $F = \langle Ar, att \rangle$ be an AF. A set $S \subset Ar$ of arguments is *conflict-free* iff there are no arguments $b, c \in S$ such that b attacks c (i.e. such that $(b, c) \in att$). F is called *strongly connected* if there is an att-path from each argument in Ar to each other argument in Ar. A *strongly connected component* (SCC) of F is a maximal strongly connected component of F. We denote the set of SCCs of F by $SCCs(F)$. When $a, b \in A$ are in the same SCC, we write $a \sim b$. Given $S \subseteq A$, we define $D_F(S) := \{b \in A \mid \exists a \in S : (a, b) \in att \wedge a \not\sim b\}$.

We now recursively define CF2 extensions as follows:

Definition 2. Let $F = \langle Ar, att \rangle$ be an AF, and let $S \subseteq Ar$. Then S is a CF2 extension of F iff either

- $|SCCs(F)| = 1$ and S is a maximal conflict-free subset of A, or
- $|SCCs(F)| > 1$ and for each $C \in SCCs(F)$, $S \cap C$ is a CF2 extension of $F|_{C - D_F(S)}$.

While the grounded extension of an AF is always unique, an AF with cycles may have multiple preferred extensions and multiple CF2 extensions. In our experiment, however, participants were asked to make a single judgment about each argument, so we compare their judgments to the *justification status* of arguments according to various semantics (see [24,4]), as the justification status is always unique for each argument. In particular, we focus on the justification statuses *strong accept* and *strong reject*, which can be defined as follows:

Definition 3. Let $F = \langle Ar, att \rangle$ be an AF, let σ be an argumentation semantics, and let $a \in A$ be an argument. We say that a is *strongly accepted with respect to* σ iff for every σ-extension E of F, $a \in E$. We say that a is *strongly rejected with respect to* σ iff for every σ-extension E of F, some $b \in E$ attacks a.

Note that in the labeling approach, strong acceptance of a corresponds to a being labeled `in` by all labelings, and strong rejection of a corresponds to a being labeled `out` by all labelings. Note that in some AFs there are arguments that are neither strongly accepted nor strongly rejected with respect to some semantics, so based on these two notions we have actually defined a three-valued partition of the arguments of a given AF with respect to a given semantics.

Let us illustrate the defined notion by considering how they apply to the three AFs depicted in Figures 1, 2 and 3, which are also the AFs used in the empirical cognitive study. In the case of the simple reinstatement AF in Figure 1, arguments C and A are strongly accepted with respect to grounded, preferred and CF2 semantics, and argument B is strongly rejected with respect to these three semantics. In the case of the floating reinstatement framework in Figure 2, argument B is strongly rejected and argument A is strongly accepted with respect to

Figure 1.
Simple reinstatement

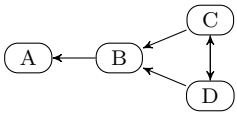

Figure 2.
Floating reinstatement

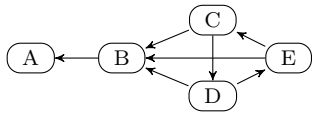

Figure 3. 3-cycle reinstatement

preferred and CF2 semantics. Arguments C and D are neither strongly accepted nor strongly rejected with respect to any of the three semantics under consideration, and with respect to grounded semantics, the same holds for arguments A and B. In the case of the 3-cycle reinstatement framework in Figure 3, none of the five arguments is strongly accepted or strongly rejected with respect to grounded or preferred semantics. With respect to CF2 semantics, on the other hand, B is strongly rejected, A is strongly accepted, while the other three arguments are neither strongly accepted nor strongly rejected.

We now briefly explain why we focus on grounded, preferred and CF2 semantics in this paper. Two other semantics widely considered in the literature are the complete semantics and the stable semantics. The justification status in the complete semantics is the same as in the grounded semantics, so we do not consider complete semantics separately. The stable semantics has the serious disadvantage that for some AFs, there is no stable extension, so the justification statuses that we defined cannot be meaningfully applied to such frameworks.

CF2 semantics belongs to a family of semantics called *naive-based* semantics, to which one also counts *naive semantics*, *stage semantics* and *stage2 semantics* (see [21]). Naive semantics ignores the directionality of attacks, a central feature that distinguishes argumentation theory from classical logic. Furthermore, no arguments ever get strongly rejected with respect to naive semantics, and only unattacked arguments get strongly accepted, so it strongly deviates from other argumentation semantics. For these reasons, whenever we make statements about "standard argumentation semantics", we consider naive semantics to be excluded from this set. Stage and stage2 semantics on the other hand are much more well-behaved and should in principle be considered as potential predictors of human argumentative reasoning. The only reason why we leave them out of the comparisons with the empirical data in Section 6 is that for the AFs used in this study, they make the same predictions as CF2 semantics, so the data from our experiment cannot distinguish between these three naive-based semantics.

2. Related Work

While formal argumentation theory is an important branch of research within AI, only a few studies have empirically investigated the cognitive plausibility of the formalisms from argumentation theory: The first of its kind was the study of Rahwan *et al.* [19], who tested how humans evaluate simple reinstatement and floating reinstatement. Cerutti *et al.* [8] have tested the correspondence between human evaluation of arguments and properties of a logic-programming-based approach to structured argumentation proposed by Prakken and Sartor [18]. Rosenfeld and Kraus [20] have empirically studied human argumentative behavior and compared it to bipolar AFs [7]. Polberg and Hunter [16] performed an experiment

to investigate the relation between human reasoning on the one hand and bipolar and probabilistic approaches to abstract argumentation on the other hand.

Of these four empirical studies, the one that comes closest in its aims to our study is the study of Rahwan *et al.* [19]. Their paper includes a discussion of why this kind of empirical validation of formalisms from argumentation theory is a highly relevant method that complements the more widely applied example-based and principle-based approaches. Being the first study to investigate the cognitive plausibility of the formalisms from argumentation theory, it laid the foundations for further work in this area. But, as is normal for an empirical study that is the first to address certain research questions, it also had some limitations and problematic features, as we explain in more detail for the rest of this section.

Their study only made use of two AFs, namely the simple reinstatement AF in Figure 1 and the floating reinstatement AF in Figure 2. Since all standard semantics agree on the evaluation of simple reinstatement, only the results on floating reinstatement could distinguish between different semantics, so only limited claims could be made about which semantics best predicts human judgments.

The authors did not empirically test their assumption that the natural language argument sets that they designed actually correspond to the intended AFs. This limitation is especially pressing in light of the fact that the attacks that they intended to be unidirectional were based on conflicts between the conclusion of the attacking argument and the premise of the attacked argument, without any indication of a preference. In the frameworks of structured argumentation from the ASPIC family [17,14,6], such undermings without preferences always give rise to bidirectional attacks.

In their study, participants were asked to assess the conclusion of a designated argument on a 7-point Likert scale from *certainly false* to *certainly true*. However, it is difficult to compare their 7-point Likert scale results to the predictions of argumentation formalisms that are two- or three-valued.

We designed our study with the aim to overcome these issues that we had identified in Rahwan *et al.*'s study, so as to make further progress based on the important foundations that they had laid with their study.

3. Cognitive variability of humans

Given that this paper presents findings of a cognitive empirical study to an audience whose scientific expertise lies mainly in areas outside of cognitive science, we present some general background from cognitive science that will help to make our methodological choices and our discussion of the results more understandable.

Humans are heterogeneous by nature; they differ from each other with respect to their cognitive abilities [1]. Cronbach [10] claimed that human heterogeneity is actually a major disturbance in the conduction of empirical studies. Cognitive variability has thus been mostly considered as an undesirable random noise in cognitive studies. This disturbance is even more problematic in the case of empirical studies that evaluate complex cognitive processes such as logical thinking and reasoning. Indeed, the inherent difficulty of such tasks not only emphasizes human differences relative to pure cognitive abilities (such as intelligence), but

also involves motivational aspects that are crucial to obtain a reliable performance from the participant [23]. In order to test the cognitive plausibility of abstract argumentation theory by minimizing unwanted bias purely related to cognition and motivation properties, we set up a methodology that favored rational thinking during the assessment.

Previous results showed that individual performance, which has generally been reported to be quite poor in pure logic and reasoning tasks, could actually be enhanced by cooperative discussion with peers. For instance, faced with the Wason selection task [22], humans solving the task in groups achieved a level of insight that was qualitatively superior to the one achieved by single individuals [13,2]. Additionally, and more generally, discussion with peers was proved to substantially improve motivation to solve a given task [15]. For these reasons, we decided to incorporate in our methodology a cooperative discussion to help participants to elaborate and enrich their thinking. This collective step with peers was designed to obtain an evaluation of the justification status more reliable than a single individual judgment. Such reliability is crucial to test the cognitive plausibility of our predictions.

4. Pilot studies

We conducted two pilot studies to verify our predictions about the existence and the directionality of attacks between natural language arguments. Both pilot studies involved the argument sets from Rahwan *et al.*'s [19] study, and new argument sets with a large variety of attack types that we specifically designed. All participants were individually assessed. In the first study, 27 undergraduate students of various backgrounds were shown two arguments at a time and were instructed to judge the acceptability of each argument (*accept*, *undecided* or *reject*). We used their judgments to infer their representation of the directionality of the attacks between the arguments. In the second study, 14 active researchers, who are expert in formal argumentation theory, had to directly judge the existence and the directionality of every attack between all arguments within each set.

Results from both pilot studies consistently showed that participants were overall able to judge the existence and directionality of a conflict between two given arguments. More particularly, the items that we designed for the pilot study led to high agreement with the theoretically assumed attack relations (74% and 84% consistent response, respectively in students and experts). On the other hand, it should be noted that attacks and non-attacks between arguments from Rahwan *et al.*'s [19] study were identified as intended in only 56% and 73% of the cases (in students and experts). Especially noteworthy is the fact that the attacks that Rahwan et al. intended to be unidirectional were identified as such by only 17% of students and 44% of experts; most participants considered them to be bidirectional rather than unidirectional attacks. The results of the pilot studies are described and discussed in more detail in a separate workshop paper [9].

For the purpose of using them in the main study, we selected four argument sets in which the agreement between the judgments of pilot study participants and the AF intended by us was especially high. The chosen arguments were judged as intended in 77% and 87% of the cases (by students and experts).

5. Design of main study

We tested 130 undergraduate students from the University of Luxembourg (mean age = 22 years). Each participant was presented with a set of 3 to 5 natural language arguments, which was designed to correspond to the simple reinstatement, floating reinstatement or 3-cycle reinstatement AF, as depicted in Figures 1, 2 and 3 in Section 1. As explained in Section 4 above, this correspondence between the argument sets chosen for the main study and the corresponding AFs had been confirmed by two pilot studies. The questionnaire consisted of two successive parts, the first of which involved drawing the attack relation between the given arguments and the second of which involved judging the acceptability of each argument in light of the information provided in all arguments. In this paper we focus our attention to the second part of the questionnaire.

The study involved arguments based on three different thematic contexts: arguments based on news reports, arguments based on scientific publications, and arguments based on the precision of a calculation tool (referred to as *mathematical context*). The full set of argument sets employed in the study is available at `http://icr.uni.lu/mcramer/downloads/Argument_Sets_COMMA_2018.pdf`. As an example, here is the argument set of the scientific context corresponding to floating reinstatement:

A. *Specimen A consists only of amylase. The 1972 Encyclopaedia of Biochemistry states that amylase is an enzyme. So specimen A consists of an enzyme.*
B. *A peer-reviewed research article by Smith et al. from 2006 presented new findings that amylase is not an enzyme. Therefore no specimen consisting only of amylase consists of an enzyme.*
C. *A study that the Biology Laboratory of Harvard University has published in 2011 corrects mistakes made in the study by Smith et al. and concludes that amylase is a biologically active enzyme.*
D. *A study that the Biochemistry Laboratory of Oxford University has published in 2011 corrects mistakes made in the study by Smith et al. and concludes that amylase is a biologically inactive enzyme.*

For every argument set {A, B, C, D} representing floating reinstatement, we used the subset {A, B, C} as an argument set representing simple reinstatement, so the data on simple reinstatement and floating reinstatement can be directly compared. Due to the particular structure of the 3-cycle reinstatement AF, we used a slightly different set of arguments for this case. Due to the difficulty of adequately designing an argument set of this complexity and verifying its correspondence to the 3-cycle reinstatement AF through pilot studies, we could only include a single argument set corresponding to the 3-cycle reinstatement AF in this study, which was of scientific context.

Participants answered the questionnaire in groups of 3 to 5 students (as a function of the number of arguments they were shown). Each argument set was used for 5 groups. Thus 45 participants responded to the simple reinstatement argument sets, 60 participants to the floating reinstatement argument set, and 25 participants to the 3-cycle reinstatement argument set.

The participants were asked to judge the acceptability of each argument from the argument set by indicating either that they *accept* the argument, that they

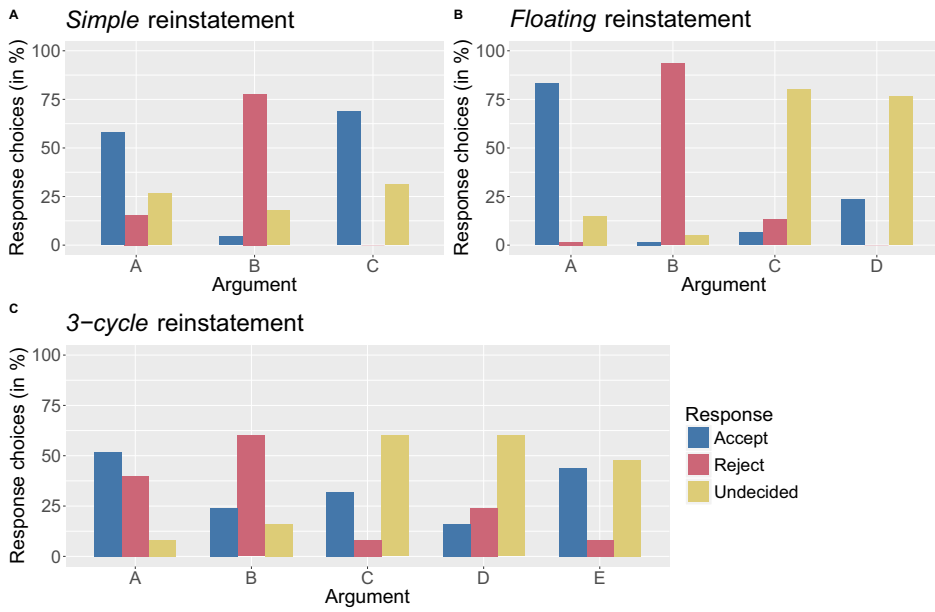

Figure 4. Results of our study disregarding the thematic context. For each each argument of each of the three AFs, the figure indicates the percentage of cases in which participants chose to *accept* a corresponding natural language argument, to *reject* it or to consider it *undecided*.

reject it, or that they consider it *undecided*. They were instructed to accept an argument unless the other arguments provide reasons to reject it. Importantly, as explained in Section 3, our methodology incorporated a group discussion to stimulate more rational thinking: In a first step, each participant had to provide a personal judgment of the acceptability of each argument. In a second step, the group members had to discuss their initial response with each other and had to reach a consensus in order to provide one response as a group (or as a function of the simple majority if no consensus was reached). Finally, in a third step, participants had the opportunity to follow or not to follow the group decision by giving their final individual response. In our presentation and discussion of the results, we only consider the results of the final individual responses.

6. Results of main study and discussion

In Figure 4, the overall results of our study are depicted, disregarding the thematic context. For simple reinstatement, a large majority of participants accepted A and C, and rejected B. A Pearson chi-squared confirmed that this distribution of the response for each argument was not due to chance, minimal $\chi^2(2) = 12.93$, $p < .001$. In sets with floating reinstatement, participants significantly tended to accept A, reject B, and consider C and D as undecided, minimal $\chi^2(2) = 55.60$, $p < .001$. Finally, for 3-cycle reinstatement, the majority of participants accepted A and rejected B while considering every other argument as undecided; their judgments reached the significance level, minimal $\chi^2(2) = 7.28$, $p = .026$.

The majority judgment about the three arguments in simple reinstatement coincides with the predictions of all standard semantics. Given that reinstatement

is one of the most fundamental features of Dung-style abstract argumentation theory, this finding speaks in favor of the cognitive plausibility of Dung-style abstract argumentation theory. Furthermore, our data is supportive of a positive answer to the first research question described in the Introduction, namely that in those cases in which all the standard argumentation semantics agree, humans evaluate the acceptability of arguments as predicted by these argumentation semantics. Of course, further AFs on which all semantics agree will have to be considered in future studies in order to test whether this really holds in this generality, or whether it only holds in limited cases.

The tendencies that we have observed above for the data that aggregates all three contexts also hold in the case of each of the three contexts considered separately, in the sense that for each argument, the response predicted by all semantics was chosen by the absolute majority ($> 50\%$) of participants, with only one exception: For argument A in the news report context, the predicted answer, *accept*, was only chosen by 40% of the agents, while 40% rejected it and 20% considered it undecided. There are two potential explanations of it: On the one hand, using world knowledge, participants might have judged the main claim in the argument (namely that Donald Trump shot a lion) to be very implausible, thus favoring to reject the argument. Additionally the difference between the contexts might be due to people trusting scientific publications and calculation tools more than they trust news reports.

In the case of floating reinstatement and 3-cycle reinstatement, there are discrepancies between the different semantics concerning the justification status of the arguments. For this reason, our results can be used to support certain answers to the second research question from the Introduction: *In those cases in which there is disagreement between the standard argumentation semantics, which one of them best predicts the judgments that humans make about the arguments?* As explained in Section 1, we focus our attention to grounded, preferred and CF2 semantics. In order to tackle this research question, we compare the predictions of these semantics pairwise by focusing on those arguments on which the two compared semantics have different justification status.

The results of these pairwise comparisons are depicted in Figure 5, separated by AF (floating reinstatement and 3-cycle reinstatement) and thematic context (as explained in Section 5 our study is limited to a single thematic context in the case of 3-cycle reinstatement). Note that the preferred and CF2 semantics fully agree on floating reinstatement, while the grounded and preferred semantics fully agree on 3-cycle reinstatement. The first two bars indicate percentage of correct predictions as a function of the two compared semantics. Since there are three possible responses (*accept*, *reject* and *undecided*), there is also always the possibility that none of the two compared semantics predicts the result correctly; the third (red) bar indicates the percentage of responses that were predicted by none of the semantics.

In floating reinstatement, the justification status varies between grounded semantics on the one hand and preferred and CF2 semantics on the other hand only in case of arguments A and B, so we focus our attention on the responses to these arguments. As can be seen on the left side of Figure 5, preferred/CF2 semantics correctly predicted the responses in 100% and in 97.5% of the cases

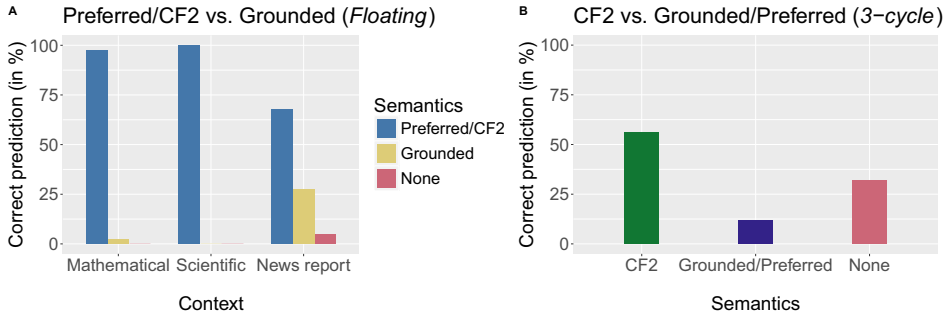

Figure 5. Pairwise comparison of semantics, aggregating all arguments whose justification status varies between the compared semantics. Data is separated by AF and thematic context. (A), on the left side, depicts predictions from the Floating reinstatement framework; (B), on the right side, depicts predictions from the 3-cycle framework. Both parts of the figure indicate percentage of cases in which participants responded as predicted by one of the compared semantics or by none of them.

in scientific and mathematical context respectively, whereas grounded semantics correctly predicted the responses to these arguments in 0% and 2.5% of the cases in these two contexts. In the case of the news report context, there is a smaller difference between the semantics, but the higher accuracy of preferred/CF2 semantics as compared to the grounded semantics is still significant (67.5% vs. 27.5%, binomial $p < .001$, one-sided). This difference between the contexts is similar to the difference between contexts discussed for simple reinstatement above, and can be explained in a similar way. So we consider it likely to be due to the influence of world knowledge on the evaluation of individual arguments rather than due to some people applying a different argumentation semantics. At any rate, note that even though the context had an influence on some people, the overall tendency that most people responded more in line with preferred and CF2 semantics than with grounded semantics holds irrespective of the context.

In 3-cycle reinstatement, we again focus on arguments A and B, as they are the only ones on which the justification status varies between grounded and preferred semantics on the one hand and CF2 semantics on the other hand. As can be seen on the right side of Figure 5, the predictions of grounded and preferred semantics were only correct in 12% of the case, whereas CF2 correctly predicted acceptance status in 56% of the cases. The difference between both semantics were significant (binomial $p < .001$, one-sided). Surprisingly, in 32% of the cases none of the semantics predicted the judgment of participants. The only explanation we have for this surprising finding is that in the 3-cycle reinstatement argument set, argument A was based on trust in a statement from the "1962 Encyclopedia of Chemistry", whereas all other scientific publications cited in this argument set were much more recent (2003 to 2008), so participants might have dismissed argument A for being based on too old a scientific source.

Summarizing the results of our comparison between the different semantics, we can say that our data strongly suggest that independently of the three thematic contexts that we have considered, CF2 and preferred semantics are a better predictor for human argument acceptance than the grounded semantics. Furthermore, the data suggest that CF2 semantics predict human argument acceptance

better than preferred semantics, but the data for this comparison is limited to a single thematic context and is less conclusive.

Note that our findings on simple reinstatement and floating reinstatement can be seen as a confirmation of the findings of Rahwan *et al.* [19] on human judgments related to these two AFs. Even though the data is not directly comparable, since their participants judged arguments on a 7-point Likert scale instead of making the three-valued acceptability judgments that our participants were asked to make, their final interpretation of their data is similar to ours: The predictions of standard semantics on simple reinstatement are in line with human judgments, and in floating reinstatement, preferred semantics predicts human judgments better than grounded semantics. We extend these findings with findings about 3-cycle reinstatement which suggest that CF2 semantics is a better predictor for human judgments than preferred semantics. But apart from this extension, there are also important methodological differences between our studies: Unlike Rahwan *et al.*, we performed pilot studies that confirmed that the natural language argument sets correspond to the intended AFs, and our pilot studies even suggest that the attacks that Rahwan *et al.* intended to be unidirectional are treated by humans as bidirectional, which calls into doubt their assumption that their findings are about the simple and floating reinstatement AFs rather than about AFs involving only bidirectional attacks. Furthermore, whereas in Rahwan *et al.'s* study participants had to respond individually, the group deliberation methodology that we have applied is known to enhance performance of humans in reasoning tasks.

7. Conclusion and Outlook

In this paper, we have presented and discussed the results of an empirical cognitive study on the cognitive plausibility of abstract argumentation semantics. Two pilot studies verified the intended link between argumentation frameworks and sets of natural language arguments. The results of the main study suggest that independently of the thematic contexts that we have considered, CF2 (or stage or stage2) and preferred semantics are a better predictor for human argument acceptance than the grounded semantics. Furthermore, the data suggest that CF2 semantics predicts human argument acceptance better than preferred semantics, but the data for this comparison is limited to a single thematic context.

This study contributes to the still young and under-explored research field of cognitive aspects of formal argumentation theory. The limitations of the presented study highlight potential further research in this field: Future studies should be based on a larger variety of AFs than the three AFs used in this study. Given that our study suggests that naive-based semantics like CF2, stage and stage2 are good predictors of human judgments, special attention should be given to AFs on which they have different predictions, so as to find out which one of them predicts human judgments best. Furthermore, future studies should attempt to minimize the influence of world knowledge on argument judgments, so as to get a clearer idea of the influence of the logical form of arguments on human judgments. One possible way in which this could be achieved is by using arguments that are embedded in a fictional setting about which the participants can have no knowledge other than the one provided to them during the experiment.

References

[1] A. Anastasi. *Differential psychology: individual and group differences in behavior*. Macmillan, 1958.

[2] M. Augustinova. Falsification cueing in collective reasoning: example of the Wason selection task. *European Journal of Social Psychology*, 38(5):770–785.

[3] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, 2011.

[4] P. Baroni, M. Caminada, and M. Giacomin. Abstract argumentation frameworks and their semantics. In P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, editors, *Handbook of Formal Argumentation*. College Publications, 2018.

[5] P. Baroni, M. Giacomin, and G. Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1):162–210, 2005.

[6] M. Caminada, S. Modgil, and N. Oren. Preferences and Unrestricted Rebut. In *Computational Models of Argument - Proceedings of COMMA 2014*, pages 209–220, 2014.

[7] C. Cayrol and M.-C. Lagasquie-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *Int. Journal of Approximate Reasoning*, 54(7):876–899, 2013.

[8] F. Cerutti, N. Tintarev, and N. Oren. Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation. In T. Schaub, G. Friedrich, and B. O'Sullivan, editors, *Proceedings of the 21st ECAI 2014*, pages 207–212, 2014.

[9] M. Cramer and M. Guillaume. Directionality of Attacks in Natural Language Argumentation. In *Proceedings of the Fourth Workshop on Bridging the Gap between Human and Automated Reasoning 2018*, in press.

[10] L. J. Cronbach. The two disciplines of scientific psychology. *American Psychologist*, 12(11):671–684, 1957.

[11] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2), 1995.

[12] W. Dvořák and S. A. Gaggl. Stage semantics and the SCC-recursive schema for argumentation semantics. *Journal of Logic and Computation*, 26(4):1149–1202, Aug 2016.

[13] D. M. M. Geil. Collaborative Reasoning: Evidence for Collective Rationality. *Thinking & Reasoning*, 4(3):231–248, 1998.

[14] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[15] J. Piaget, L. Smith, T. Brown, R. Campbell, N. Emler, and D. Ferrari. *Sociological Studies*. Routledge, 1995.

[16] S. Polberg and A. Hunter. Empirical evaluation of abstract argumentation: Supporting the need for bipolar and probabilistic approaches. *Int. Journal of Approximate Reasoning*, 93:487–543, 2018.

[17] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[18] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7(1-2):25–75, 1997.

[19] I. Rahwan, M. I. Madakkatel, J.-F. Bonnefon, R. N. Awan, and S. Abdallah. Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.

[20] A. Rosenfeld and S. Kraus. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems*, 6(4):30:1–30:33, 2016.

[21] L. van der Torre and S. Vesic. The principle-based approach to abstract argumentation semantics. In P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, editors, *Handbook of Formal Argumentation*. College Publications, 2018.

[22] P. C. Wason. Reasoning. In B. Foss, editor, *New Horizons in Psychology*, pages 135–151. Harmondsworth: Penguin Books, 1966.

[23] B. Weiner. *Theories of motivation; from mechanism to cognition*. Markham psychology series. Markham Pub. Co., 1972.

[24] Y. Wu and M. Caminada. A Labelling-Based Justification Status of Arguments. *Studies in Logic*, 3(4):12–29, 2010.