Natural Language Engineering (2023), **29**, pp. 1150–1187
doi:10.1017/S1351324923000062

SURVEY PAPER

Argumentation models and their use in corpus annotation: Practice, prospects, and challenges

Henrique Lopes Cardoso[1,*], Rui Sousa-Silva[2], Paula Carvalho[3] and Bruno Martins[3,4]

[1]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC/LASI), Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, [2]Centro de Linguística da Universidade do Porto (CLUP), Faculdade de Letras da Universidade do Porto, Via Panorâmica, 4150-564 Porto, Portugal, [3]INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal, and [4]Instituto Superior Técnico (IST), Av. Rovisco Pais, 1049-001 Lisboa, Portugal
*Corresponding author. E-mail: hlc@fe.up.pt

(Received 3 August 2022; revised 28 January 2023; accepted 30 January 2023; first published online 28 February 2023)

Abstract

The study of argumentation is transversal to several research domains, from philosophy to linguistics, from the law to computer science and artificial intelligence. In discourse analysis, several distinct models have been proposed to harness argumentation, each with a different focus or aim. To analyze the use of argumentation in natural language, several corpora annotation efforts have been carried out, with a more or less explicit grounding on one of such theoretical argumentation models. In fact, given the recent growing interest in argument mining applications, argument-annotated corpora are crucial to train machine learning models in a supervised way. However, the proliferation of such corpora has led to a wide disparity in the granularity of the argument annotations employed. In this paper, we review the most relevant theoretical argumentation models, after which we survey argument annotation projects closely following those theoretical models. We also highlight the main simplifications that are often introduced in practice. Furthermore, we glimpse other annotation efforts that are not so theoretically grounded but instead follow a shallower approach. It turns out that most argument annotation projects make their own assumptions and simplifications, both in terms of the textual genre they focus on and in terms of adapting the adopted theoretical argumentation model for their own agenda. Issues of compatibility among argument-annotated corpora are discussed by looking at the problem from a syntactical, semantic, and practical perspective. Finally, we discuss current and prospective applications of models that take advantage of argument-annotated corpora.

Keywords: Argumentation models; Corpus annotation; Argument mining

1. Introduction

The study of argumentation, that is, the linguistic and rhetorical strategies used to justify or refute a standpoint, with the aim of securing agreement in views (van Eemeren, Jackson, and Jacobs 2015), dates back to ancient times, with Aristotle being often acknowledged as one of the seminal scholars devoted to this intrinsically human activity. A vast amount of work on the study of argument and argumentation processes has been conducted over the centuries, with a recent treatment of both classical and modern backgrounds appearing in van Eemeren *et al.* (2014). Besides its usefulness in fields such as philosophy, linguistics, and the law, argumentation has also gained attention in computer science, with a long tradition in artificial intelligence research. Within this field, argumentation has been studied in knowledge representation and reasoning (including non-monotonic or defeasible reasoning Nute 1994; Pollock 1995) and also as a means to develop sophisticated interactions (e.g., in expert systems Ye and Johnson 1995, or to perform automated negotiation in multi-agent systems Rahwan *et al.* 2003). Several studies have been developed on abstract argumentation frameworks, including the seminal work by Dung (1995). More recently, argumentation has been explored in the context of computational linguistics and natural language processing (NLP) (Lippi and Torroni 2016; Moens 2018; Lawrence and Reed 2019).

Several approaches to model argumentation have been proposed, with different degrees of expressiveness. An often-used distinction in the study of argumentation considers two perspectives (O'Keefe 1977; Reed and Walton 2003): *argument as product* and *argument as process*. The former is more focused on studying how arguments are internally structured, while the latter looks at how arguments are used in the context of dialogical interactions (regardless of whether the interlocutor is explicitly present or not). Analyzing argument structure amounts to looking at its logic and deals with issues related to the reasoning steps employed and their validity. Studying arguments as ingredients of an argumentation process borrows a dialogical (in fact dialectical) conceptualization, including both verbal (e.g., live debates) and written communication (e.g., manifestos or argumentative essays). Modeling such a process entails considering the persuasive nature of argumentation, or at least its influencing character (Huber and Snider 2006). Even though this distinction between product and process is not clear-cut, most of the proposed theoretical argumentation models can be seen as relying mostly on one of these perspectives. For instance, Freeman's *standard approach* (Freeman 2011) (after the work of Thomas 1986) focuses on the relations between argument components (premises and claims) and thus primarily studies argument *as product*.[a] At the other end of the spectrum, models that take into account argumentative interactions, for example, by anticipating possible attacks (such as in Toulmin's model Toulmin 1958) or by explicitly handling dialogues (such as in Inference Anchoring Theory Reed and Budzynska 2011) clearly approach arguments *as process*. Several other models have been proposed and framed within a different taxonomy (Bentahar *et al.* 2010): *monological*, *dialogical*, and *rhetorical* models of argumentation. We review the most relevant proposals in Section 2.

Within artificial intelligence, the use of computational models of natural argument is particularly relevant in the flourishing area of *argument mining* (Stede and Schneider 2018), which aims at automatically extracting argument structures from natural language text. A combination of NLP techniques and machine learning (ML) methods has been used for this purpose. Although approaches to unsupervised language representations and knowledge transfer between different NLP tasks are on the agenda, argument mining is an extremely demanding task in terms of semantics. As such, any successful approach will have to rely on annotated data (Pustejovsky and Stubbs 2012) to apply supervised learning methods. In fact, several argument annotation projects have been put in place, following different argumentation models, focusing on different text genres, and having different aims. We will analyze the most relevant projects in Section 3. As in all of NLP tasks, most available corpora are written in English.

This is, in effect, the main research question we address: *How have theoretical argumentation models been used in corpus annotation?* While doing so, we critically explore, in Section 4, issues of compatibility between the output of argument annotation projects, also in light of current trends in the field of NLP and machine learning.

Thus, this article makes the following contributions. Firstly, we provide an updated account of existing argumentation models and highlight those that can be or have already been explored in corpus annotation. Secondly, we present a survey on existing argument annotation projects, taking the identified argumentation models as a starting point. Finally, we discuss the challenges and prospects of combining existing argument-annotated corpora. We end this survey by prospectively looking at the usefulness of argument-annotated corpora in several foreseeable applications, some of which are already being explored.

Despite the existence of somewhat recent surveys in this domain (namely, Lippi and Torroni 2016; Budzynska and Villata 2018; Cabrio and Villata 2018; Lawrence and Reed 2019;

[a] Freeman does use the term *macrostructure of arguments* (Freeman 1991), but the distinction between micro- and macrostructure has been set out differently (Bentahar, Moulin, and Bélanger 2010).

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Schaefer and Stede 2021), most of them are task-oriented, focus more on argument mining techniques, or do not put in perspective the usage of theoretical argumentation models for building annotated corpora. On the contrary, our starting point is to review argumentation models and to look at how they have been used and adapted with the aim of analyzing argumentative discourse and ultimately producing annotated corpora for various tasks.

2. Argumentation models

Given the wide study of argumentation in different fields—such as philosophy, linguistics, the law, and artificial intelligence—unsurprisingly several distinct models have been proposed to handle argumentation from different perspectives. In the classical Aristotelian theory, persuasion can be argumentatively explored in three dimensions: *logos*, *pathos*, and *ethos*. Logos is concerned with the construction of logical arguments, pathos focuses on appealing to the emotions of the target audience, and ethos explores the credibility of the arguer or of an entity mentioned in the discourse. Similarly, philosophers have long established a distinction between logical, dialectical, and rhetorical argumentation perspectives (Blair 2003). A different, albeit similar, taxonomy has been put forward by Bentahar *et al.* (2010), who proposed a classification into monological, dialogical, and rhetorical argumentation models. The similarity comes from the fact that monological models focus on the links between argument components, for example, how conclusions are drawn from premises, thereby providing an account for logical connections. Moreover, the dialectical/dialogical distinction (Wegerif 2008) is rather subtle. In essence, it focuses on the contents of the expressed views in an interaction: while dialectics assumes the presence of opposing or conflicting views that are in need of settlement, dialogical interactions are not necessarily argumentative. In any case, they share this spirit in light of the proposed taxonomy. The rhetorical perspective focuses on the nature of argumentation as a means of persuasion and thus takes into account the audience's perception of the employed arguments. Instead of looking primarily at the structure of arguments, rhetorical models are more concerned with how a discourse's target audience is convinced or affected by it.

After discussing, in Section 2.1, the nature of argumentative text, in Section 2.2 we present existing argumentation models, with a bias towards those that can be or have been explored in corpus annotation. This will provide the needed grounds to better appreciate the annotation projects that we will later discuss.

2.1. Argumentative text

Argumentation can take various forms, either spoken, written, or graphical; it can even be multimodal by taking on a combination of these. In argumentation, it is implicitly assumed that someone, a so-called *reasonable critic* (van Eemeren 2001), will take in the expressed arguments and possibly be convinced by them. Yet, the immediate presence of an interlocutor is what distinguishes a dialogue from a monologue, the latter being *one of the most prolific forms of persuasive argumentation* (Reed and Long 1997). Going beyond dyadic relationships, a complex multi-party discussion has been framed into the notion of a polylogue (Kerbrat-Orecchioni 2004; Lewiński and Aakhus 2014).

Regardless of the immediate availability of the interlocutor of an argumentation exercise, different kinds of argumentative texts exist, whose differences we explore by connecting argumentative discourse with text genres.

2.1.1. Variations in argumentative content

A genre is a communicative event characterized by a set of communicative purposes (Bhatia 1993) and extra-textual elements (Biber 1988), whose external form and use situations determine how

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

each text is perceived, categorized, and used by the members of a community (Swales 1990). Texts of the argumentative type can take the form, for example, of essays, debates, opinion articles, or research articles: they are all argumentative in the sense that they aim to persuade the other party (Hargie, Dickson, and Tourish 2004), but take different communicative functions and hence are of different genres. Consequently, depending on the particular genre, they can be more or less argumentatively structured.

Explicit argumentative texts (such as persuasive essays) are highly structured and usually include a thesis-antithesis-synthesis type of structure (Schnitker and Emmons 2013) while employing linguistic argumentative devices. Other less structured genres (such as opinion articles) tend to express arguments in a less orderly manner: argumentation tends to be more subtle and resorts to enthymemes, in which an argument includes implicit and omitted elements. Furthermore, argumentative (discourse) connectors (van Eemeren, Houtlosser, and Snoeck Henkemans 2007; Das and Taboada 2013, 2018) are often avoided, and the logical reasoning is left to the reader to establish. When argumentative discourse markers or connectors are prevalent (such as in legal texts or persuasive essays), identifying arguments becomes a comparatively easier task; conversely, when such linguistic clues are omitted or mostly absent (such as in user-generated content or in opinion articles, considered a free and fluid text genre), a careful interpretation of the text content is needed to identify argumentative reasoning steps.

It is thus pertinent to analyze different text genres in terms of argumentative content. This is a crucial aspect that should be considered when implementing an argument annotation project.

2.1.2. Argumentation and text genres

Argumentation in *debates* (O'Neill, Laycock, and Scales 1925) has a long tradition in human societies and is particularly impactful in public visibility settings, such as election debates (Haddadan, Cabrio, and Villata 2018; Visser *et al.* 2020) or policy regulations (Lewiński and Mohammed 2019). Debates are a (typically oral) form of interaction where participants express their views on a common topic. They enable two or more participants to respond to previously expressed arguments, for example, by attacking them or providing alternative points of view, either on the fly or in a turn-based fashion. Debates are usually highly argumentative in nature, and, in most cases, each participant aims to convince the audience that their standpoint prevails. Exploring argumentative strategies in debates is no longer exclusive to humans: as a recent demonstration (Aharonov and Slonim 2019) has shown that machines too can use them, although perhaps not with the competence and versatility with which humans can employ such strategies. This raises new challenges and concerns about the ethical usage of machines to conduct debates.

Persuasive essays and *legal texts* (especially court decisions) are among the most argumentative text genres. Essays (Burstein *et al.* 1998) and legal texts (both in case Wyner *et al.* 2010 and civil law do Carmo 2012 documents) are typically well-structured in terms of discourse and make use of explicit markers and connectors that point to the presence of argumentative reasoning structures. The field of legal reasoning is actively exploring ways to automate the handling of legal documents through computational means (Rissland 1988; Eliot 2021).

Speeches and *manifestos*, in particular those with a political bias (Menini *et al.* 2018, 2017), may involve content as structured as essays. However, some speeches include a mix of informational and persuasive elements, and thus, argumentative structures may not be as prevalent as in persuasive essays or legal texts.

Scientific articles (Gilbert 1976; Lauscher, Glavaš, and Eckert 2018a) should have, at least in principle, a markedly scientific writing style, including sections such as introduction, related work, proposal, experimental evaluation, discussion, or conclusions. Their argumentative structure often follows a predictable pattern (Teufel 1998; Wagemans 2016b), which can be exploited to look for arguments in specific sections.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

News editorials and *opinion articles* (Bal and Saint Dizier 2010) typically have a loose structure while at the same time including relevant argumentative material. In these texts, however, argumentation structures do not necessarily abide by standard reasoning steps and often include figures of speech such as metaphors, irony, or satire. This more or less free writing style makes it harder to clearly identify arguments in text, since the lack of explicit markers hinders the task of telling descriptive from argumentative content.

Review articles such as product reviews often include a mix of description and comparative evaluation. As such, the employed argument schemes are expected to be rather focused (Wyner *et al.* 2012). Finally, certain kinds of *user-generated content*, including opinions, customer feedback, complaints, comments (Park and Cardie 2014), and social media posts (Goudas *et al.* 2014; Schaefer and Stede 2021), often raise a number of challenges, related to poorly substantiated claims or careless writing. This kind of content is also often short, lacking enough context to allow for a deep understanding of the author's arguments, when available. Hence, it is a more difficult genre to process from an argumentation point of view.

The diversity of argumentative content across different text genres must be considered both when drafting annotation guidelines and when developing argument mining models. More specifically, analyzing texts in terms of argumentative content may justify the adoption of particular argumentation models. We next survey such models, while in Section 3 we analyze annotation efforts that have been carried out, targeting different text genres.

2.2. Theoretical models of argumentation

According to van Eemeren *et al.*, the general objective of argumentation theory is *to provide adequate instruments for analyzing, evaluating, and producing argumentative discourse* (van Eemeren *et al.* 2014, p. 12). There are now several attempts to provide such instruments, with varying degrees of complexity. In this section, we review some of the most relevant ones. We highlight those that can be or have been used in corpus annotation. We will contrast those with a few others that are useful for more qualitative argument analysis.

Arguably, the most basic model that can be used for argumentation is propositional logic (Govier 2010, ch. 8). In fact, deductive argumentation (Besnard and Hunter 2001) can be represented through propositions and propositional logic connectives (*and*, *or*, *not*, *entailment*). However, classical (propositional) logic is not appropriate to deal with conflicting information (Besnard and Hunter 2008, pp. 16-17), which is prevalent in argumentative reasoning.

2.2.1. Toulmin's model

One of the most influential argumentation models of the 20th century is Toulmin's (Toulmin 1958). The model is composed of six argument components (as illustrated in Figure 1), which together articulate a theoretically sound and well-formed argument: a *claim*, whose strength is marked by a *qualifier*, is supported by certain assumptions in the form of *data (grounds)*; this support is based on a *warrant* (a general and often implicit, commonsense rule whose applicability is explainable by a *backing*), unless certain circumstances, stated in the form of a *rebuttal*, occur. An example is included in Figure 2.

Despite its apparent comprehensiveness, Toulmin's model has been criticized for its lack of applicability to real-life arguments. Freeman (1991) and Simosi (2003) agree that the model is more adequate for analyzing arguments *as process* (in a dialectical setting) than *as product*. In fact, the model takes for granted that certain elements, such as data, warrant, or backing, are put in place in response to critical questions raised by a challenger. However, this is seldom the case in everyday arguments. Furthermore, according to Freeman, the distinction between these elements is not clear, making it hard to identify them when analyzing argumentative texts.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

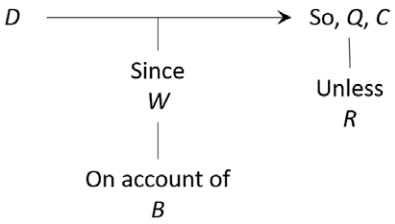

Figure 1. Toulmin model, adapted from Toulmin (1958). *D* = *Data*, *Q* = *Qualifier*, *C* = *Claim*, *W* = *Warrant*, *B* = *Backing*, *R* = *Rebuttal*.

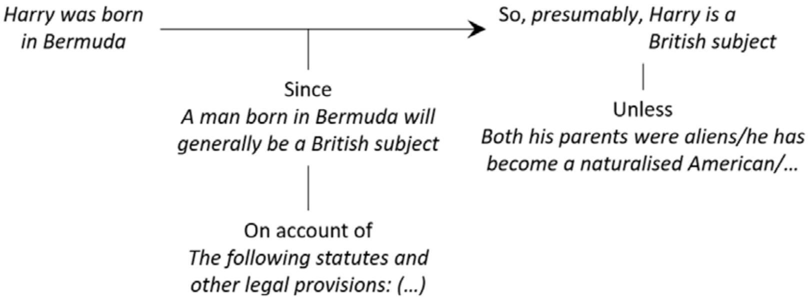
Harry was born in Bermuda
So, presumably, Harry is a British subject
Since
A man born in Bermuda will generally be a British subject
Unless
Both his parents were aliens/he has become a naturalised American/...
On account of
The following statutes and other legal provisions: (...)

Figure 2. An instantiation of Toulmin's model, adapted from Toulmin (1958).

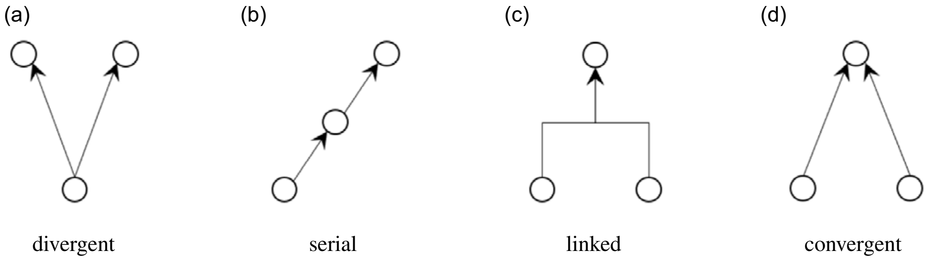

Figure 3. Argument structures, adapted from Freeman (2011).

2.2.2. Freeman's model

By building upon the criticisms of Toulmin's model, and starting from Thomas' work (Thomas 1986) (dubbed as the *standard approach*), Freeman (1991; 2011) proposes to represent argument *as product* using diagrams that include as main elements simply *premises* and *conclusions*. These fit together by forming various structures (illustrated in Figure 3): *divergent*, where a premise is given to support two distinct conclusions; *serial*, where a chain of premise, intermediate conclusion, and final conclusion is used; *linked*, where two premises are needed to support, together, a conclusion; and *convergent*, where a conclusion is independently supported by two different premises. These structures can be combined to obtain argument structures of arbitrary complexity.

According to Freeman, the notion of warrant and backing in Toulmin's model is something that is typically not explicit in an argument *as product* but rather elicited by a challenger. Given its interpretation as an inference rule, a warrant fits nicely as an element of *modus ponens* and is thus representable as a premise in a linked structure. In his proposal of an *extended standard approach*, Freeman introduces a reinterpretation of the qualifier, which he renames *modality*. Instead of

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

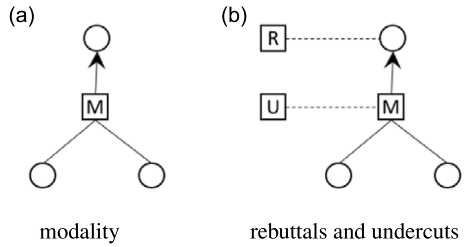

Figure 4. Extended standard approach, adapted from Freeman (2011).

qualifying the conclusion, a modality indicates how strongly the combination of the premises supports the conclusion. The example (from Freeman 2011, p. 18) “All humans are mortal. Socrates is a human. So, *necessarily*, Socrates is mortal” includes a linked argument structure with two premises, where “necessarily” plays the role of the modality—it does not qualify the conclusion, but rather the strength of the argumentative reasoning step. Another example (adapted from Freeman 2011, p. 18) is “That die came up one in the previous 100 tosses. Therefore, it is *probable* that it will come up one on the next toss,” where “probable” is the modality. Diagrammatically, Freeman suggests representing modalities as labeled nodes that are interposed between premises and conclusion (see Figure 4a).

To refine Toulmin’s notion of rebuttal, Freeman borrows Pollock’s distinction between *rebutting* and *undercutting* defeaters (Pollock 1995) (i.e., *countermoves*). While a rebutting defeater may present evidence that the conclusion is false, an undercutting defeater questions the reliability of the employed inferential step from premises to conclusion. Freeman suggests drawing both kinds of defeaters similarly to Toulmin’s approach, connecting them to the modality. In Figure 4b, however, we distinguish the cases for undercutting and rebutting defeaters in a way that more clearly illustrates their different targets, which is in line with Pollock’s approach (Freeman found this distinction in notation to be unnecessarily complicated). Finally, Freeman introduces counters to defeaters (both rebutting and undercutting), which thus indirectly support the conclusion.

2.2.3. Walton’s argumentation schemes

To harness the diversity of argumentation used in practice, several taxonomies have been developed over the years (Kienpointner 1986, 1992; Pollock 1995; Walton 1996; Katzav and Reed 2004; Lumer 2011). These so-called *argumentation schemes* (Macagno, Walton, and Reed 2017) vary in depth and coverage and attempt to capture stereotypical patterns of inference occurring in argumentation practice. They thus depart from logic-based deductive forms of reasoning, focusing instead on how humans express themselves in natural language.

The concept of argumentation scheme is so prevalent that many other theoretical models of argumentation encompass their own schemes, such as different functions of warrant in Toulmin’s model (Toulmin 1958), or argument classifications in the New Rhetoric (Perelman and Olbrechts-Tyteca 1969). Of the existing taxonomies, however, Walton’s argumentation schemes (Walton 1996) are arguably one of the most well-known and widely used, even though it has gone through several variations (Walton, Reed, and Macagno 2008; Walton and Macagno 2015).

While other more elaborated argumentation theories define arguments functionally, a theory focusing on argumentation schemes empirically collects and classifies arguments as used in actual practice, distinguishing them by their contents (following mostly a bottom-up approach). Aggregating arguments according to their similarities gives rise to elaborate taxonomies, containing more than sixty different schemes (Walton *et al.* 2008), from which one may identify the main categories. Walton *et al.* (2008) identify three— *reasoning arguments*, *source-based arguments*, and *arguments applying rules to cases*— while Walton and Macagno (2015) divide the first of these

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

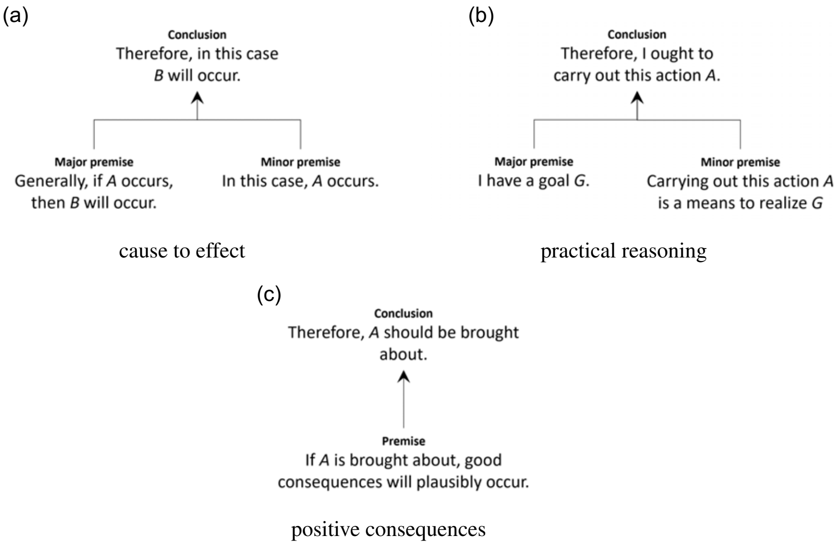
(a)
Conclusion
Therefore, in this case B will occur.
Major premise
Generally, if A occurs, then B will occur.
Minor premise
In this case, A occurs.
cause to effect
(b)
Conclusion
Therefore, I ought to carry out this action A.
Major premise
I have a goal G.
Minor premise
Carrying out this action A is a means to realize G
practical reasoning
(c)
Conclusion
Therefore, A should be brought about.
Premise
If A is brought about, good consequences will plausibly occur.
positive consequences

Figure 5. Argument scheme examples, adapted from Macagno *et al*. (2017).

main categories into *discovery arguments* and *practical reasoning*. Figure 5 shows some examples of argument schemes.

Besides its own pattern of premises and conclusion, in Walton's approach, each argumentation scheme encompasses a critical perspective on the strength of an argument through a set of *critical questions*. These represent the defeasibility conditions of the argument or the weaknesses it may suffer from and may be used to search for attack clues. Examples of critical questions regarding the argument scheme from practical reasoning illustrated in Figure 5b include (Macagno *et al.* 2017):

CQ1 What other goals do I have that might conflict with *G*?

CQ2 What alternative actions to *A* would also bring about *G*?

CQ3 Among *A* and these alternative actions, which is arguably the most efficient?

CQ4 What grounds are there for arguing that it is practically possible for me to do *A*?

CQ5 What consequences of doing *A* should also be taken into account?

Critical questions may be relied upon to build a strong (and virtually unbeatable) argument or used by an attacker to refute the argument—which grants the model a dialogical flavor.

2.2.4. Argumentum model of topics

One alternative proposal for argument schemes, focusing on the inferential steps of argumentation, has been proposed by Rigotti and Greco under the name of *Argumentum Model of Topics (AMT)* (Rigotti and Greco Morasso 2010; Rigotti and Greco 2019). This model distinguishes between two components of argument schemes: the *procedural* component and the *material* component.

The procedural component focuses on the semantic-ontological structure generating the inferential connection that is the basis for the argument's logical form. Within it, three levels are identified: the *locus* concerns the ontological relation on which argumentative reasoning is based; each locus gives rise to *inferential connections* called *maxims*; finally, maxims activate *logical forms*,

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

such as *modus ponens* or *modus tollens*. An example provided by Rigotti and Greco Morasso (2010, pp. 499-500) goes as follows: "(1) It is true of this evening (our national holiday) that there will be traffic jams. (2) Because the fact that there were traffic jams was true for New Year's Eve. (3) And the national holiday is comparable to New Year's Eve." Here, the semantic-ontological relation is one of locus from analogy by comparing the "national holiday" to "New Year's Eve." The inferential connection is that if something ("traffic jams") has been the case for a circumstance ("New Year's Eve") of the same functional genus as X ("national holiday"), it may be the case for X as well. Finally, the logical form employed is *modus ponens*, when instantiating the inferential connection with "traffic jams," "New Year's Eve" and "national holiday."

The material component provides applicability of the inferential connection to an actual argument by exploring both implicit and explicit premises. For the above example, the fact that "the national holiday is comparable to New Year's Eve" requires an effectual backing coming from an outside material starting point shared by the interlocutors. AMT claims to make it possible to explain and reconstruct the inferential mechanism employed in argumentative discourse, deemed to be often complex or implicit.

In terms of argument schemes (or *loci*), AMT proposes a taxonomy based on three domains (Rigotti and Greco Morasso 2009; Musi and Aakhus 2018). The *syntagmatic/intrinsic* domain concerns things that are referred to in the standpoint and aspects ontologically linked to it, including arguments from definition, from extensional implications, from causes, from implications and concomitances, and from correlates. The *paradigmatic/extrinsic* domain refers to arguments based on paradigmatic relations, including those from opposition, analogy, alternatives, or termination. Finally, *middle/complex loci* lie in the intersection between intrinsic and extrinsic ones and include arguments from authority, from promising and warning, from conjugates, and from derivates.

2.2.5. Inference Anchoring Theory

Specifically exploring the dialogical nature of argumentation, Reed and Budzynska propose the *Inference Anchoring Theory (IAT)* (Budzynska and Reed 2011; Reed and Budzynska 2011) as a means to offer an explanation of argumentative conduct in terms of the anchoring of reasoning structures in persuasive dialogical interactions. In this sense, IAT provides a bridge between the reasoning (logical) and the communicative (dialogical) dimensions of argumentation.

IAT relies on the fact that the logical links between dialogical utterances are governed by dialogue rules (transitions) expressing how sequences of utterances can be composed (Budzynska *et al*. 2014). These transitions act as "anchors" to the underlying reasoning steps employed in the dialogue. The propositions and relations that together form the argumentative reasoning are anchored, by means of illocutionary connections, in the locutions and transitions that constitute the dialogue. IAT includes *locutions* (dialogical units) and *transitions* between them; *illocutions* connect the communicative and the reasoning dimensions, particularly by connecting locutions to *propositions*; propositional *relations* recover the reasoning that is anchored in the discourse. Different relations between propositions can be used, distinguishing not only between inference, conflict, or rephrase, but also encompassing argumentation schemes.

Figure 6 shows the general structure of IAT usage: the flow of the dialogue is depicted, on the right side of the figure, through (usually chronologically ordered) transitions; propositions and reasoning steps are shown on the left side anchored in illocutions in the discourse. Examples of illocutions (or illocutionary connections) include *asserting*, *questioning*, *challenging*, *conceding*, *(dis)agreeing*, *restating*, or *arguing*; some of them are illustrated in the example shown in Figure 7.

2.2.6. Periodic Table of Aarguments

The wide variety of taxonomies of argumentation schemes hinders their adoption in different fields, including artificial intelligence (Katzav and Reed 2004). Wagemans (2016a) points out

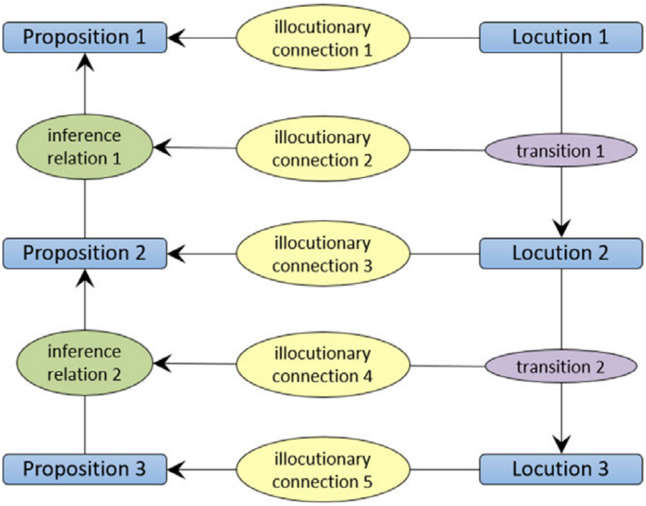

Figure 6. IAT model.

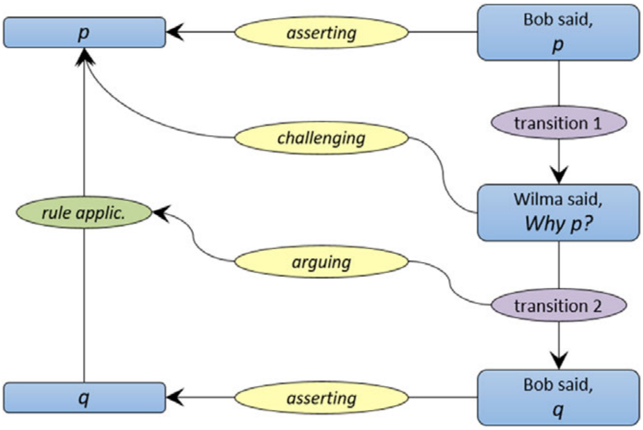

Figure 7. An instantiation of the IAT model, adapted from Budzynska *et al*. (2014): an assertion p is challenged and then justified using q, resulting in the inference relation $q \implies p$.

that existing theories are not based on a formal ordering principle, making them harder to use. Because they often follow a bottom-up approach, no theoretical rationale is provided regarding the number of categories identified.

To circumvent these problems, Wagemans proposes a *Periodic Table of Arguments (PTA)* (Wagemans 2016a), which orthogonally classifies arguments in three dimensions. The first dimension approaches the propositional content of both premises and standpoints (conclusions) to distinguish between *predicate* and *subject* arguments: if both premise and standpoint share the same subject, we are in the presence of a predicate argument, as the arguer aims to increase the acceptability of the standpoint by exploiting a relationship (through a common subject) between predicates; conversely, if the argument elements share the same predicate, we have a subject argument, as the arguer exploits a relationship (through a common predicate) between subjects. The second dimension identifies *second-order*, as opposed to *first-order*, arguments. The former is the case when someone else's standpoint—an assertion—is used to reconstruct another standpoint, as

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

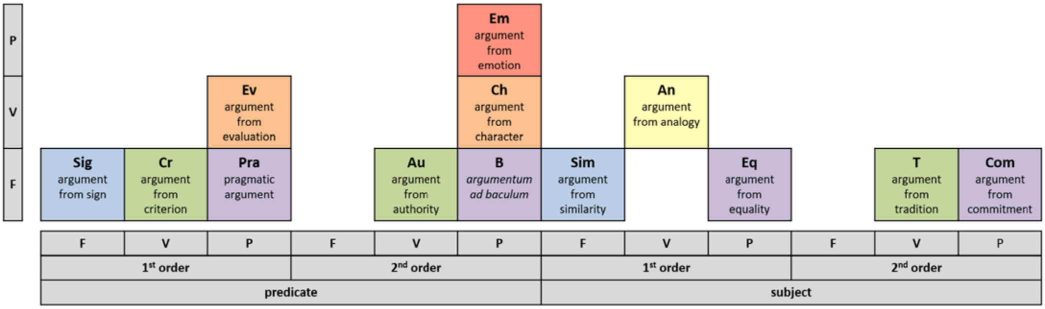
P
V
F
Em argument from emotion
Ev argument from evaluation
Ch argument from character
An argument from analogy
Sig argument from sign
Cr argument from criterion
Pra pragmatic argument
Au argument from authority
B argumentum ad baculum
Sim argument from similarity
Eq argument from equality
T argument from tradition
Com argument from commitment
F V P F V P F V P F V P
1st order
2nd order
1st order
2nd order
predicate
subject

Figure 8. Periodic table of arguments, adapted from Wagemans (2016a).

happens in the scheme argument from expert opinion. First-order arguments do not rely on such kind of assertion but work rather on propositions (Wagemans 2019). Finally, the third dimension details the types of propositions included in both premise and standpoint, and their combinations. A proposition of *policy (P)* expresses a specific policy that should be carried out. A proposition of *value (V)* expresses a judgment towards an arbitrary subject. A proposition of *fact (F)* denotes an assertive observation about the subject. The argument is characterized by the pairs of types of propositions employed.

When combined, these three dimensions give rise to 36 different argument types (some of which are shown in Figure 8), which can be represented as an acronym capturing the classification of the argument in each dimension. For instance, *1 pre VF* stands for a first-order (*1*) predicate (*pre*) argument connecting a proposition of fact premise to a proposition of value conclusion. An example would be "unauthorized copying is not a form of theft, since it does not deprive the owner of use" (Wagemans 2016a). As another example, "you should take his medicine, because it will prevent you from getting ill" (Wagemans 2016a) is a first-order predicate argument combining a proposition of fact premise with a proposition of policy conclusion—*1 pre PF*.

2.2.7. Other models

The argumentation models listed so far have been used in corpora annotation projects, which will be discussed in Section 3. A few other models are worth mentioning, given their importance in the analysis of argumentation practice. These models have been used for qualitatively analyzing arguments and for building argumentation support systems.

The *New Rhetoric* (Perelman and Olbrechts-Tyteca 1958, 1969) was proposed by Perelman and Olbrechts-Tyteca as an effort to recover the Aristotelian tradition of argumentation and is focused on rhetoric as a means of persuasion. In this sense, the targeted audience is a central issue in this model, as argumentation strategies are used to convince or persuade others. The New Rhetoric focuses on the analysis of rational arguments in "non-analytic thinking" (i.e., reasoning based on discursive means with the aim of persuasion)—it does not attempt to prescribe ways to practice rational argumentation, but rather to describe various kinds of argumentation that can be successful in practice. Perelman and Olbrechts-Tyteca define the new rhetoric as "the study of the discursive techniques allowing us to induce or to increase the mind's adherence to the theses presented for its assent" (van Eemeren *et al.* 2014, p. 262). For such techniques to be successful, the first step is to understand who is the audience of the discourse. Differently from formal logic-based approaches, argumentation is considered sound if it is effective in convincing the target audience. Perelman and Olbrechts-Tyteca describe a taxonomy of argument schemes where premises are divided into two classes: those accepted by a universal audience (e.g., facts) and those that deal with preferences. The latter are thus subjective in nature and can be used effectively if they match the preferences (values and value hierarchies) of the audience. The schemes are grouped into two

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

distinct processes: *association* (connecting elements conceived as separate by the audience) and *dissociation* (splitting something seen as a whole into separate elements).

Developed by van Eemeren and Grootendorst, *pragma-dialectics* (van Eemeren and Grootendorst 2004; van Eemeren 2018) is an argumentation theory that aims to analyze and evaluate argumentation in actual practice. Instead of focusing on argument *as product* or *as process*, pragma-dialectics looks at argumentation as a discursive activity, that is, as a complex speech act (van Eemeren and Grootendorst 1984) with specific communicative goals. According to pragma-dialectical theory, argumentation is part of an explicit or implicit dialogue, in which one participant aims at convincing the other on the acceptability of their standpoint. When the dialogue between the writer and prospective readers is implicit, the former must elaborate by foreseeing any doubts or criticisms that the latter may have or bring about. The kind of analysis proposed by pragma-dialectics is based on a dialog model, on which a qualitative review is employed, for instance, to detect fallacious arguments (van Eemeren and Grootendorst 1987; Visser, Budzynska, and Reed 2017). In terms of argument structure, pragma-dialectics distinguishes between *multiple*, *subordinatively compound*, and *coordinatively compound* argumentation. As a way of comparison, these resemble, respectively, the convergent, linked, and serial structures in the Freeman model (Freeman 1991, 2011).

After discussing the most relevant theoretical models of argumentation, in the next section we focus on the usage of such models in argument annotation projects.

3. Argument annotation

The increase in research on argument mining has brought a need to build annotated corpora that provide enough data for supervised machine learning approaches to succeed in this domain. Several annotation projects have been undertaken, based on different argumentation models, such as those described in Section 2.2.

We address the topic of conducting argument annotation projects in Section 3.1. Then, we survey some of the most relevant projects that closely follow a specific argumentation model in Section 3.2. In Section 3.3, we discuss other annotation projects which do not explicitly adopt any of such models.

3.1. Annotation effort

Any manual corpus annotation task is very time-consuming and requires an appropriate level of expertise (although some annotation projects rely exclusively on crowdsourcing-based approaches Snow *et al.* 2008; Rodrigues, Pereira, and Ribeiro 2013; Nguyen *et al.* 2017; Habernal and Gurevych 2016b; Stab *et al.* 2018). When dealing with argumentation, this requirement is considerably more critical, given the intricate process of detecting arguments in written text. The implicit or explicit nature of arguments is largely determined by the genre of the text, as discussed in Section 2.1. Some text genres are therefore more susceptible to different human interpretations, due to their free structure and writing style; depending on the genre of the target corpus, identifying arguments properly can be subjective and context-dependent, often reliant on background and world knowledge (Lauscher *et al.* 2022). The demanding nature of an argument annotation project is determined by two interdependent main axes: (i) the nature or genre of the target corpus and (ii) the argumentation theory employed. Typically, argumentatively rich and explicit text sources are less difficult to annotate. In this kind of text, arguments are often expressed by employing strong discourse markers, thus making it easier to interpret and extract the reasoning adopted by the author. This explains why most annotation projects rely on such kinds of sources, including persuasive essays (Stab and Gurevych 2014; Musi, Ghosh, and Muresan 2016) or legal texts (Mochales and Ieven 2009; Zhang, Nulty, and Lillis 2022).

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

The argumentation model to adopt is largely related to the respective text genre. When deciding on a specific argumentation model to follow, there is often a trade-off between expressiveness and annotation complexity. Some models, while providing a rich set of building blocks that may be useful to identify and fully characterize arguments within the text, are more demanding for the annotator. This entails the need for a more elaborated set of annotation guidelines, together with careful training. Furthermore, harder annotation tasks that make use of complex annotation schemes bring two main concerns. On the one hand, there is an aggravated need for expert annotators, caused by increasing cognitive demands. On the other hand, high inter-annotator agreement is less likely to occur when using larger annotation schemes (Bayerl and Paul 2011), due to annotator bias (Lumley and McNamara 1995). Such personal bias is worsened when using a larger set of annotators, as it may translate into different interpretations of annotation guidelines. For these reasons, annotation projects often fall back to a shallow approach, which in some cases allows for exploring crowdsourcing (Habernal and Gurevych 2016b; Nguyen *et al.* 2017; Stab *et al.* 2018).

3.2. Annotation projects grounded in argumentation models

The study of argumentation from a philosophical and linguistic perspective has given rise to a number of studies on its use in practice. Encouraged by the application of NLP in the analysis of written arguments and also by a flourishing research interest in argument mining (Stede and Schneider 2018), a number of projects have been conducted to build linguistic resources annotated with arguments—argument corpora. There is considerable variability in the theoretical grounding of the existing argument annotation projects. While some of them try to employ a specific argumentation model, in many cases the model is simplified to accommodate the characteristics of the corpus or the project aims. However, building a corpus on the basis of a short-term project's goal may limit its potential use in future research. As was argued by Reed *et al.* (2008), one of the key roles that a corpus can play is providing a foundation for multiple projects.

Table 1 shows a collection of argument annotation efforts that have produced corpora by following, to a significant extent, a theoretical model of argumentation.[b] To collect this list, we searched for published works in LREC,[c] in top-tier NLP conferences (including *ACL, EMNLP, and COLING), in the Argument Mining workshop series (Al-Khatib, Hou, and Stede 2021), and also those cited in available surveys (Lippi and Torroni 2016; Lawrence and Reed 2019). Furthermore, we favored those for which the produced corpus has been made publicly available.

In several cases, simplifications or adaptations are made to the underlying theoretical argumentation model. We now discuss these annotation projects in further detail.

3.2.1. Toulmin annotations

An attempt to use Toulmin's model to annotate arguments has been made by Habernal and Gurevych (2017). This project targeted a dataset of English user-generated content and newswire articles about six controversial topics in education: homeschooling, public versus private schools, redshirting, prayer in schools, single-sex education, and mainstreaming. In total, 340 documents have been annotated.[d]

The authors proposed a modified Toulmin model, which includes some of the original elements— *claim*, *backing*, *rebuttal* —and adds *premise* and *refutation*. These changes are based on a number of observations: the lack of explicit mentions to cogency degrees in the analyzed texts, which hinders the detection of the *qualifier*; the fact that *warrants* are often implicit and fail to be found in practice; the need to consider rebuttal attacks, dubbed *refutations*, which are

[b] For convenience, Table 1 is ordered alphabetically by argumentation model.

[c] http://www.lrec-conf.org/.

[d] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2423.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Table 1. Annotated corpora following argumentation models

Model	Corpus	Genre	Language	Size
Argumentum Model of Topics (AMT)	Musi *et al*. (2016)	Persuasive essays	English	30 essays
	Argumentative Microtext Corpus (Musi *et al*. 2018)	Short argumentative texts	English	112 texts
Freeman	Argument-Annotated Essays (Stab and Gurevych 2014, 2017; Eger *et al*. 2018)	Persuasive essays	English, German	402 essays
	Argumentative Microtext (Part 1) (Peldszus and Stede 2016)	Short argumentative texts	German, English	112 texts
	Argumentative Microtext (Part 2) (Skeppstedt, Peldszus, and Stede 2018)	Short argumentative texts	English	171 texts
	Rocha *et al*. (2022b)	Opinion articles	Portuguese	373 articles
Inference Anchoring Theory (IAT)	Mock mediation (Janier and Reed 2017)	Dispute mediation	English	1 document
	US2016 (Visser *et al*. 2020)	TV political debate and Social media	English	3 debates + live reactions
	QT30 (Hautli-Janisz *et al*. 2022)	Broadcast political debate	English	30 episodes
Periodic Table of Arguments (PTA)	US2016 G1tvWAGEMANS (Visser *et al*. 2018)	TV political debate	English	1 debate
Toulmin	Argument-Annotated User-Generated Web Discourse (Habernal and Gurevych 2017)	User-generated content, Newswire articles	English	340 documents
Walton's argumentation schemes	AraucariaDB (Reed *et al*. 2008)	News editorials, Parliamentary records, Judicial summaries, Discussion boards	English	+700 analyses
	(Mochales and Ieven 2009)	Legal texts	English	45 judgments
	(Hansen and Walton 2013)	Political argumentation in news media	English	233 reports
	(Reisert *et al*. 2017a,b)	Short argumentative texts	English	89 texts
	US2016 G1tvWALTON (Visser *et al*. 2018)	TV political debate	English	1 debate

used to ensure the consistency of the argument's position; the equivalence between *grounds* and *premises*, the latter being used in several works in argument mining; the reinterpretation of *backing* as additional, non-essential evidence of support (often a stated fact) to the whole argument, as opposed to a support to the *warrant*, which is left out. Furthermore, claims are allowed to be implicit, and annotators have been asked to put forward the stance of the author in such cases. In this new setting, refutation looks similar to Freeman's notion of counters to countermoves applied to the warrant of the argument, that is, counter rebutting/undercutting defeaters (Freeman 2011, pp. 24-25). An example of application of this modified Toulmin model is shown in Figure 9.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

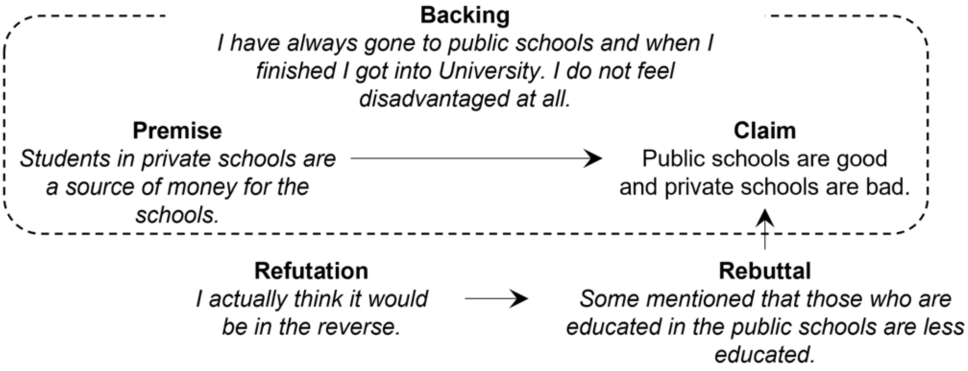
Backing
I have always gone to public schools and when I finished I got into University. I do not feel disadvantaged at all.
Premise
Students in private schools are a source of money for the schools.
Claim
Public schools are good and private schools are bad.
Refutation
I actually think it would be in the reverse.
Rebuttal
Some mentioned that those who are educated in the public schools are less educated.

Figure 9. Modified Toulmin model example, adapted from Habernal and Gurevych (2017).

The authors report a moderate inter-annotator agreement. Unsurprisingly, they have found that a source of disagreement between annotators was the easy confusion between refutations and premises, as they both function as support for the claim. The (modified) Toulmin model has been found to be better suited for short persuasive documents, and more problematic in the longer ones explored in the project. This is because, in many cases, the discourse presented in such documents has been found to be more rhetoric than argumentative – argument components are not employed with specific argumentative functions in mind in the *logos* dimension (the one focused on by the Toulmin model).

3.2.2. Freeman annotations

Given its more simple nature based on premises and conclusions with different kinds of support and attack relations, Freeman's argumentation model has been used in some annotation projects, although in many cases without exploiting its full potential. Peldszus and Stede (2016) conducted a controlled text generation experiment of 112 short argumentative texts, followed by their annotation—the Argumentative Microtext Corpus.[e] Each text was triggered by a specific question, and it should be a short, self-contained sequence of about five sentences providing a standpoint and justification. All texts were originally written in German and have been professionally translated into English. After validating the developed annotation guidelines, the final markup of argumentation structures in the full corpus was done by one expert annotator. The annotation scheme, further detailed in Peldszus and Stede (2013), includes *linked* and *convergent* arguments, as well as two types of attacks: *undercuts* and *rebuttals*. Furthermore, the special case of *support by example* is considered.

Skeppstedt *et al.* (2018) have extended this original work, exploring crowdsourcing as a means to increase the size of the short text corpus. Each participant was asked to write a 5-sentence long argumentative text on a particular topic, in English, while making its stance clear through a claim statement and providing at least one argument for the opposite view. The resulting 171 texts have then been annotated (in a non-crowdsourced way) following the same guidelines. An example is shown in Figure 10. The main difficulties reported in the annotation process include implicit claims, restatements, direct versus indirect supports, the distinction between argument support and mere causal connections, implicit annotator evaluations, and the existence of non-argumentative text units.

[e] http://angcl.ling.uni-potsdam.de/resources/argmicro.html.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

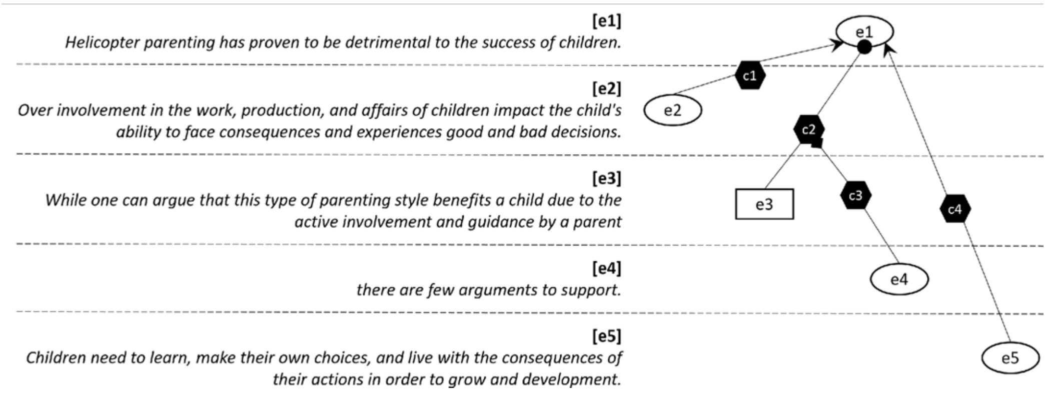
[e1]
Helicopter parenting has proven to be detrimental to the success of children.
[e2]
Over involvement in the work, production, and affairs of children impact the child's ability to face consequences and experiences good and bad decisions.
[e3]
While one can argue that this type of parenting style benefits a child due to the active involvement and guidance by a parent
[e4]
there are few arguments to support.
[e5]
Children need to learn, make their own choices, and live with the consequences of their actions in order to grow and development.
e1
c1
e2
c2
e3
c3
c4
e4
e5

Figure 10. Freeman example annotation in the Argumentative Microtext Corpus, adapted from Skeppstedt *et al.* (2018). Rounded elements support the central claim (e1), while the boxed element critically questions it; arrow-headed connections are supports, the circle-headed is a rebuttal, and the square-headed is an undercut.

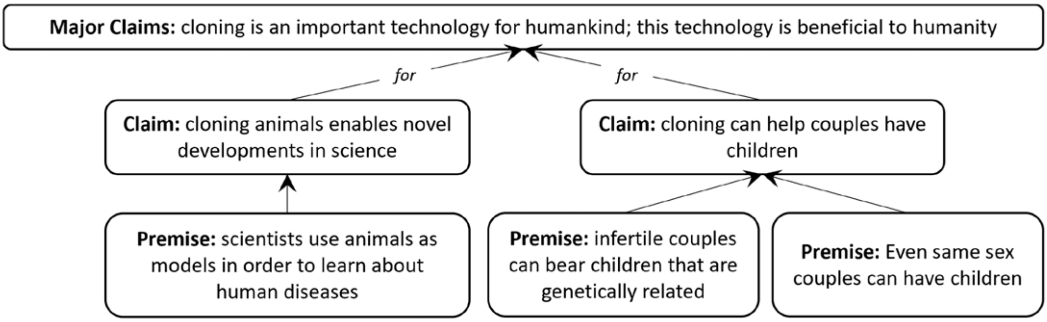
Major Claims: cloning is an important technology for humankind; this technology is beneficial to humanity
for
for
Claim: cloning animals enables novel developments in science
Claim: cloning can help couples have children
Premise: scientists use animals as models in order to learn about human diseases
Premise: infertile couples can bear children that are genetically related
Premise: Even same sex couples can have children

Figure 11. Freeman example annotation in the Argument-Annotated Essays Corpus, adapted from Stab and Gurevych (2017).

Stab and Gurevych (2014) worked on a particularly rich argumentative text genre—persuasive essays — to create an annotated corpus of 90 documents[f] where the argumentative structure of each document includes a *major claim*, *claims*, and *premises*. A major claim expresses the author's standpoint concerning the topic and is assumed to be expressed in the first paragraph (introduction) and possibly reinstated in the last one (conclusion). Each argument is composed of a claim and at least one premise. Claims are labeled with a stance attribute (*for* or *against* the major claim). Two kinds of directed relations between premises and claims are included: *support* and *attack*. Both relations can hold between a premise and another premise and between a premise and a claim. An example is shown in Figure 11.

Some simplifications to the Freeman model have been introduced by Stab and Gurevych based on the analysis of the argumentative essay corpus contents. The annotation process did not consider the linked versus convergent structures distinction, namely because it has been considered ambiguous (Freeman 2011, p. 91). Also, divergent arguments have been excluded from consideration to model the argumentation used in the essays with tree structures. The authors report inter-annotator agreements for argument component annotations, for the stance attribute, and

[f] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2421.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

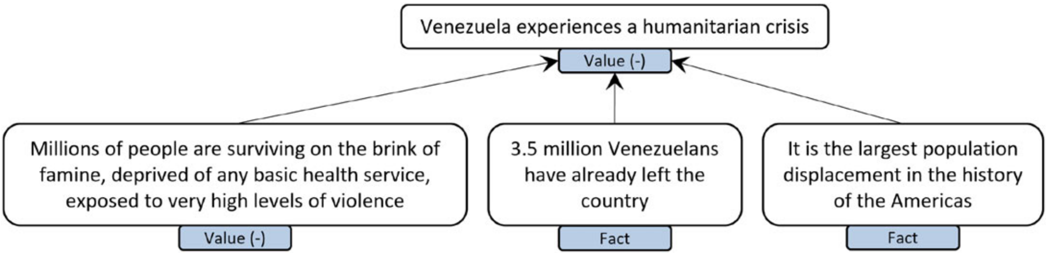
Venezuela experiences a humanitarian crisis
Value (-)
Millions of people are surviving on the brink of famine, deprived of any basic health service, exposed to very high levels of violence
Value (-)
3.5 million Venezuelans have already left the country
Fact
It is the largest population displacement in the history of the Americas
Fact

Figure 12. Freeman example annotation in the opinion articles corpus, adapted (and translated) from Rocha *et al.* (2022b).

argumentative relations. The most significant source of confusion among annotators is the distinction between claims and premises, originated by the fact that chains of reasoning can be established (serial structures Freeman 2011).

Stab and Gurevych (2017) have extended this work, this time creating an annotated corpus of 402 English documents,[g] with similar inter-annotator analysis. Eger *et al.* (2018) provide a fully parallel human-translated English-German version of this corpus and machine-translate the corpus into Spanish, French, and Chinese. In both cases, argumentation annotations are kept, either manually or automatically projected from the original annotated corpus.

Rocha *et al.* (2022b) have made some inroads into applying Freeman's model to a less structured argumentative genre—opinion articles as published in a Portuguese newspaper. In their approach, each of a set of 373 articles has been annotated considering the full Freeman model, including several kinds of argument structures (linked, convergent, divergent, and serial) as well as two kinds of relations (support and attack). As an additional annotation layer, each annotated proposition (seen as a premise or a conclusion) was classified into one of *fact*, *policy*, or *value* (in this case distinguishing between positive, negative, and neutral), closely following part of the Periodic Table of Arguments schema. An example annotation is shown in Figure 12.

Given the expected difficulty in annotating arguments in this kind of text, they opted to have three independent annotations per document. Arguments are assumed to be contained in a single paragraph. The authors carry out an extensive inter-annotator agreement analysis, in which they have found that token-level agreement on argumentative units is challenging. However, for the argumentative units in which there is agreement, higher-level component analysis (such as types and roles of propositions, and macro-structure of arguments) can obtain considerable agreement. Based on this corpus, they have explored the recent trend into perspectivist approaches to NLP (Basile *et al.* 2021), by considering different ways of aggregating annotations (Rocha *et al.* 2022a).

3.2.3. Walton's argumentation schemes annotations

Walton's argumentation schemes have been used, to a different extent, in several argumentation studies. In terms of annotated corpora, some projects are worth being mentioned. One of the first efforts in analyzing and annotating argumentative structures in text is part of the Araucaria project (Rowe and Reed 2008). The AraucariaDB corpus[h] comprises a set of argumentative examples extracted from diverse sources and geographical regions. The source material, in English, includes newspaper editorials, parliamentary records, judicial summaries, and discussion boards.

Work on the corpus started in 2003. Reed *et al.* (2008) provide a concise description of the development principles behind the construction of this corpus and discuss potential uses of this kind of language resources. Araucaria is based on different Walton argumentative scheme sets.

[g] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422.
[h] http://corpora.aifdb.org/araucaria.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Reed *et al.* concede that argument annotation is a subjective task, and for that reason, no careful inter-coder agreement analysis has been carried out on Araucaria.

Mochales and Ieven (2009) explore legal cases from the European Court of Human Rights (ECHR), in order to analyze how judges of this court present their arguments. In this kind of legal case, an applicant presents a complaint about the violation of a specific article of the European Convention on Human Rights. A corpus of 45 judgments and decisions written in English has been annotated. The documents follow a very well-defined structure, and the part where the court's argumentation is expressed is clearly identifiable.

The authors have characterized arguments by distinguishing discourse relations as per pragma-dialectics (*multiple*, *coordinative*, and *subordinative* relations) and further classify arguments following Walton's argumentation schemes. Starting from Walton (1996), but with a mixture of scheme types and subtypes, Mochales and Ieven proposed a list of 26 schemes, of which 6 have been found to be more frequent: argument *from analogy*, *from established rule*, *from consequences*, *from sign*, *from precedent*, and *from example*, in decreasing order of occurrences. A simplified version of the corpus, including premises, conclusions, and their relations, is available as the ECHR Corpus (Poudyal *et al.* 2020).[i]

The most particular cause of disagreement between annotators was the fact that the ECHR corpus, or more specifically the document sections that were targeted, include "reported arguments", where the court revises the arguments employed by both the plaintiff and the defendant. Some annotators regarded this as argumentation, while others did not. Furthermore, disagreement on distinguishing arguments from facts has also been observed, an issue related to the fact that annotators had diverse legal backgrounds, bringing different conceptions of law. Another major source of disagreement is related to argument structure. Given the predominance of the complexly structured argument *from established rule* scheme in the ECHR corpus (losing only to argument *from analogy*), considerable disagreement has been observed due to premises/conclusions being identified as conclusions/premises, and with subordinative structures being mistakenly identified as coordinative. Finally, distinguishing between sub-classes of argumentation schemes has been found to be much harder than keeping with more general classes.

Hansen and Walton (2013) have studied political public interventions in the context of the 2011 Ontario provincial election. Their data source is indirect: they analyze the politicians' arguments as reported in four Canadian newspapers, which they have monitored during the election period, in order to collect reported arguments that could be attributed to a candidate. In total, they have collected 256 distinct argument events. They initially followed a subset of 14 argumentation schemes, taken from Walton (2006, pp. 132-137). Hansen and Walton point out that they do not assume that either politicians or reporters have any knowledge of argument schemes, which are seen as conceptual tools that interest analysts, not necessarily argument makers. After finding that some of the arguments analyzed did not fit any of the schemes in the list and that a couple of schemes were not applied to any arguments, the authors changed the list to one with 21 schemes, where 9 new schemes were added.

Starting with the Argumentative Microtext Corpus, (Peldszus and Stede 2016), Reisert *et al.* (2017a,b) have applied an additional annotation layer based on Walton's argumentation schemes (Walton *et al.* 2008). A total of 89 texts from that corpus were annotated with several rhetorical patterns, covering argumentative relations of support, rebuttal, and undercut. Such patterns have been created mainly for the *argument from consequence* scheme.

Visser *et al.* (2018) analyze TV political debates on the US 2016 presidential election. This dataset had already been annotated with IAT (to be discussed later), but in this work, the authors have followed two additional theories to annotate the inference relations identified: Walton's argumentation schemes and Wagemans' Periodic Table of Arguments. In terms of the former, the authors make use of an extensive list of 60 schemes (the main schemes from Walton *et al.* 2008), to

[i] http://www.di.uevora.pt/pq/echr/.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

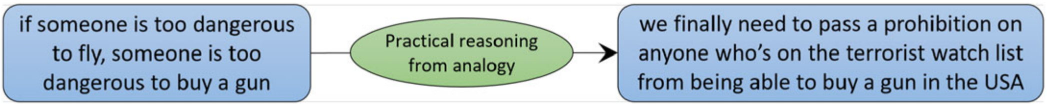
if someone is too dangerous to fly, someone is too dangerous to buy a gun
Practical reasoning from analogy
we finally need to pass a prohibition on anyone who's on the terrorist watch list from being able to buy a gun in the USA

Figure 13. Walton's argumentation scheme example annotation in the US 2016 presidential election Corpus, adapted from Visser *et al*. (2018).

annotate the first general election debate between Hilary Clinton (Democrat) and Donald Trump (Republican).[j] Two annotators were employed, who used a classification decision tree prepared by the authors—an intuitive means of distinguishing between the schemes, taken to be mutually exclusive. An example is shown in Figure 13. In practice, only 39 different schemes have been identified, with a predominance of *argument from example*. Visser *et al.* also point out that one of the most highlighted academic schemes—*argument from expert opinion* —is quite rare, which is illustrative of the discrepancies between a scholarly take on argument schemes and their practical prevalence. As expected, some classes of argument schemes turned out to be particularly difficult to distinguish, such as *practical reasoning* and *argument from values*. Only 14 of the original 505 inference nodes have been marked with a *default inference* label, meaning that the annotators did not manage to frame them within one of the 60 schemes. Substantial agreement between annotators is reported.

3.2.4. Argumentum Model of Topics annotations

The Argumentum Model of Topics has been comparatively less used in practice (also due to the fact that it is a much more recent proposal). Here we briefly cover two works by Musi *et al.* on applying AMT to Argument-Annotated Essays (Musi *et al*. 2016)[k] and to the Argumentative Microtext Corpus (Musi *et al*. 2018).[l]

Musi *et al*. (2016) developed a pilot annotation study of 30 persuasive essays. The authors hypothesize that AMT has the potential to enhance the recognition of argument schemes, by arguing that it offers, unlike other theoretical models, a taxonomic hierarchy based on distinctive and mutually exclusive criteria. Such criteria appeal to semantic properties that are part of premises and claims, rather than the logical forms of arguments. However, annotations following AMT require reconstruction of implicit premises: common ground knowledge, inferential rules, or intermediary claims.

Working on top of the Argument-Annotated Essays corpus (Stab and Gurevych 2014), the authors base their annotation project on first identifying the argument scheme, from either *intrinsic* (definitional, mereological, causal), *extrinsic* (analogy, opposition, practical evaluation, alternatives), or *complex* (authority) relations. For that, a set of identification questions and linguistic clues was provided. Only support relations were considered. Annotators were then asked to identify the *inferential rule* at work (for which representative rules for each argument scheme were provided). This annotation exercise has come to a slight agreement outcome, with a considerable degree of confusion between the employed schemes.

Following the previous work and using an updated set of guidelines, Musi *et al.* (2018) applied AMT to the Argumentative Microtext Corpus (Peldszus and Stede 2016), by annotating its 112 short texts. When doing so, both support and rebut relations were considered, again making use of the eight middle-level schemes of AMT and the associated inference rules. The annotation process has achieved only fair agreement, again due to confusion between some of the employed schemes (most notably argument from practical evaluation and the default "no argument").

[j] http://corpora.aifdb.org/US2016G1tvWALTON.
[k] https://github.com/musielena/argscheme_aclworkshop2016.
[l] http://angcl.ling.uni-potsdam.de/resources/argmicro.html.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

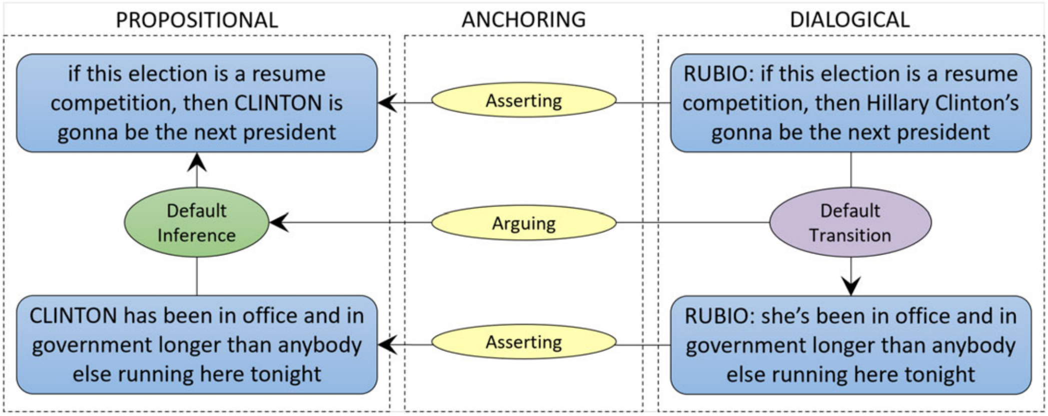
PROPOSITIONAL
ANCHORING
DIALOGICAL
if this election is a resume competition, then CLINTON is gonna be the next president
Asserting
RUBIO: if this election is a resume competition, then Hillary Clinton's gonna be the next president
Default Inference
Arguing
Default Transition
CLINTON has been in office and in government longer than anybody else running here tonight
Asserting
RUBIO: she's been in office and in government longer than anybody else running here tonight

Figure 14. IAT example annotation in the US 2016 presidential election Corpus, adapted from Visser *et al*. (2020).

3.2.5. Inference Anchoring Theory annotations

The more elaborate nature of Inference Anchoring Theory, whose full power is harnessed when analyzing dialogical interactions, has been explored in recent annotation efforts. Janier and Reed (2017) focus on dispute mediation discourse. The aim of dispute mediation is to help conflicting parties in finding a way to solve their dispute by resorting to a mediator third party. Given the usually confidential nature of such disputes, Janier and Reed resort to a 45-page transcript document of a mock mediation session provided by Dundee's Early Dispute Resolution service.[m] According to the authors, the transcription is taken to be realistic and useful for the purpose of studying how dialogues unfold in the particular context of dispute mediation—how mediators suggest arguments and deal with impasses, how the arguments exchanged between conflicting parties form a reasonable discussion, and so on. Janier and Reed annotate the document using IAT schemes,[n] thereby exposing the "shape of the discussion," that is to say, the argumentative structure of the dialog. They then analyze whether mediation-specific argumentative moves can be easily detected and how they configure mediation tactics and strategies.

Visser *et al.* (2020) describe a publicly available corpus of argumentatively annotated debate, which makes use of a detailed approach to argument analysis, using IAT. They analyze TV political debates on the US 2016 presidential election, together with reactions to the debates on social media. More specifically, the authors have focused on the first Republican and the first Democrat primary debates and on the first general election debate between the candidates of both parties (Donald Trump and Hillary Clinton). An example annotation is shown in Figure 14. Besides analyzing the debates themselves, Visser *et al.* have also collected online live (i.e., contemporaneous) reactions on Reddit. By annotating inter-textual correspondence, they provide an unusually rich corpus.[o] The inclusion, in their study, of this social media debate thread brings greater diversity in language use, given the variability of participants' backgrounds and even nationalities. Still, Visser *et al.* observe that Reddit discussions also contain well-structured argumentative content.

To validate the annotation process and compute inter-annotator agreement, approximately 10% of the corpus (word count) was annotated by two independent annotators. When combined and normalized by overall corpus word count, the annotations of the various sub-corpora obtained substantial agreement. When analyzing each annotation sub-task (including segmentation, transitions, illocutionary connections, and propositional relations), the authors observed that the annotation of illocutionary connections was more challenging than the other sub-tasks,

[m] https://www.dundee.ac.uk/edr.
[n] http://corpora.aifdb.org/mockmediation.
[o] http://www.corpora.aifdb.org/US2016.

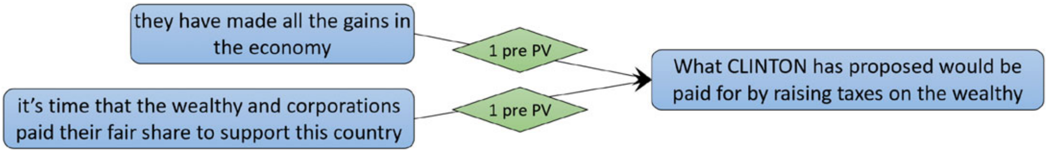
they have made all the gains in the economy
1 pre PV
it's time that the wealthy and corporations paid their fair share to support this country
1 pre PV
What CLINTON has proposed would be paid for by raising taxes on the wealthy

Figure 15. PTA example annotation in the US 2016 presidential election Corpus, adapted from Visser *et al*. (2018).

due to the closeness between certain types of connections. They also observed higher agreement for the Reddit subcorpus than for the TV debates. Reasons for this include the shorter dialogue turns and the explicit response structure between posts in the thread, making it easier to identify discourse transitions and relations. In terms of propositional relations, those of inference have been found to be predominant in both the TV debates and Reddit sub-corpora. Moreover, conflict relations were more frequent in the Reddit discussion, while rephrases and reformulations were more identified in the TV debates.

Intertextual correspondence explored the fact that some Reddit participants draw conclusions based on arguments presented in the TV debates or show disagreement by first rephrasing what the candidates have said. This annotation layer has created new transitions, illocutionary connections, and propositional relations between the two genres of sub-corpora. For obvious reasons, in the explored scenario Reddit participants could react to the actual candidates' utterances in the debate, but not vice versa. Visser *et al.* observe that rephrase relations are the most common, indicating the expected practice that Reddit contributors often restate what the candidates have said during the debate, in order to support their own online arguments.

More recently, Hautli-Janisz *et al.* (2022) have explored using IAT in broadcast political debate, focusing on 30 episodes of BBC's "Question Time" from 2020 and 2021. The authors claim that the obtained corpus, QT30,[p] is the largest corpus of broadcast debate and report moderate agreement for the annotations. While they make use of nine illocutionary connections, they have found that more than half of those annotated correspond to assertions (which is typical in dialogical argumentation). Hautli-Janisz *et al.* also distinguish between three roles in debates: *moderator*, *panel*, and *audience*. They analyze whether patterns of conflict or support differ between roles and verify that most attacks are observed between panel members (as expected) while a much lesser number of supports between different speakers is obtained.

3.2.6. Periodic Table of Arguments annotations

We are aware of a single annotation project making use of the Periodic Table of Arguments model. As mentioned earlier, Visser *et al.* (2018) analyze TV political debates on the US 2016 presidential election. They do so by extending the original IAT-based inference relation annotations with Walton's argumentation schemes and with Wagemans' model, focusing on the first general election debate between Hilary Clinton and Donald Trump. The Periodic Table of Arguments model has been applied to its full extent, focusing on (i) inference relations so as to classify them as *first-order* or *second-order*, and as *predicate* or *subject* arguments, and (ii) the corresponding information nodes (sources and targets of relations), classifying them as propositions of *fact*, *value*, or *policy*. An example is shown in Figure 15.

Based on a sample of approximately 10% of the corpus annotated by two different annotators, the annotation process revealed substantial to almost perfect agreement, combining the three partial classifications (the lowest partial agreement concerns the distinction between first and second-order arguments and is attributed to its unbalanced nature, with an overwhelming

[p] http://corpora.aifdb.org/qt30.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

predominance of first-order arguments). From the generated corpus,[q] Visser *et al.* obtain a final coding of each argument with one of the 36 possible types in the Periodic Table of Arguments. Failure to classify an argument in one of the partial classification tasks has led to a *default inference* fallback, which turned out to represent a significant number of cases (approximately 17%, much more than that observed with Walton's argumentation schemes for the same corpus, as reported in Section 3.2.3). This is, however, not surprising since the argument schemes result directly from the classification of both inference relations and the intervening information nodes—they are not chosen by the annotator from a list, as with Walton's argumentation schemes.

3.3. Other annotation efforts

It is worth noting that most annotation projects focus on the *logos* dimension of the classical Aristotelian distinction, or at least do not explicitly consider the *pathos* and *ethos* dimensions. Exceptions to this include Duthie *et al.* (2016), who explore the ethos dimension, Habernal and Gurevych (2017), who consider the pathos alongside the logos dimension, and Hidey *et al.* (2017), who consider annotating premises with one of the three dimensions of logos, pathos, and ethos. Several other argument annotation projects have been carried out without a clear adherence to one of the argumentation models introduced in Section 2. We refer to some of those works here.

Aharoni *et al.* (2014) annotate 586 Wikipedia articles on 33 controversial topics, following a simple claim-evidence structure. For each document, a *context-dependent claim* is identified, together with *context-dependent evidence*, which are further classified into one of *study*, *expert*, or *anecdotal*. The corpus includes a total of 2683 argument elements, including 1392 claims and 1291 evidences.

In the domain of political debates, Haddadan, Cabrio, and Villata (2019) perform a manual argument mining effort targeting 39 political debates from 50 years of US presidential campaign debates. The output is a corpus of 29k argument components, labeled as premises and claims, but without any relation links between them. The authors observe that the corpus contains a higher number of claims as compared to premises, which is explained by the fact that political candidates often put forward arguments without providing premises for their claims.

Focusing on news editorials, Al-Khatib *et al.* (2016) aim at mining argumentation strategies. According to the authors, editorials lack a clear argumentative structure and frequently resort to enthymemes. After employing an automatic segmentation of editorials into argument discourse units (ADU), annotators were asked to annotate each ADU according to one of the following roles: *common ground*, indicating common knowledge or a generally accepted truth; *assumption*, when the unit states an assumption or opinion of the author, or a general observation or (possibly false) fact; *testimony*, stating a proposition made by some expert, authority, witness, and so on; *statistics*, when expressing quantitative evidence; *anecdote*, transmitting a personal experience of the author or a specific event; and *other*, when the unit does not add to the argumentative discourse or does not match any of the previous roles.

While there seems to be some connection between such roles and some of Walton's argumentation scheme elements, the authors' purpose is to analyze argumentation strategy at a macro-document level, as opposed to a finer granularity of analyzing individual arguments. The corpus obtained[r] is composed of 300 editorials from *Al Jazeera*, *Fox News*, and *The Guardian*. While analyzing the corpus, the authors observed the highest proportion of *assumptions* in *The Guardian*, with *Fox News* strongly relying on *common ground* and having twice as many *testimony* evidence when compared to *The Guardian*; *Al Jazeera* emphasizes *anecdotal* discourse units. All three sources resort to *statistics* in a similar way.

[q] http://corpora.aifdb.org/US2016G1tvWAGEMANS.
[r] https://webis.de/data/webis-editorials-16.html.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

The Internet Argument Corpus (Walker *et al.* 2012) includes annotations of debate on internet forums obtained in a crowdsourcing setting. Covering 10 different topics, the chosen annotation scheme was based on several dialogic and argumentative markers (namely related to degrees of agreement, emotionality, cordiality, sarcasm, target, and question nature), assigned to quote-response pairs and to chains of three posts. The second version of the corpus (Abbott *et al.* 2016)[s] contains entries from three distinct online fora, with a total of 482k posts.

Habernal and Gurevych have also addressed argumentation in debate portals with a rather shallow approach. They have cast the problem to a relation annotation task, in which crowdsourced annotators have been asked to select the most convincing argument from a pair (Habernal and Gurevych 2016b)[t] In the sequel, they have extended the corpus with reason annotations (Habernal and Gurevych 2016a),[u] for which they have used a hierarchical annotation process guided by questions leading to one of 19 distinct labels.

To understand what makes a message persuasive, Hidey *et al.* (2017) have explored annotating comments in the popular Reddit ChangeMyView online persuasion forum. They follow a two-tiered crowdsourced annotation scheme, where premises are labeled according to their persuasive mode (ethos, logos, or pathos), while claims are labeled as interpretation, evaluation, agreement, or disagreement—which also capture the dialogical nature of the corpus. Egawa, Morio, and Fujita (2019) have worked on the same online forum and use five types of elementary units— *fact*, *testimony*, *value*, *policy*, and *rhetorical statement* —and two types of relations— *support* or *attack*. The resulting corpus contains 4612 elementary units and 2713 relations in 345 posts.

A shallower approach to dealing with debate datasets is followed by Durmus and Cardie (2018), who collected debates from different topic categories together with votes from the readers of the debates. In parallel, they collected user information (political and religious beliefs) intending to study the role of prior beliefs when assessing debate winners. They have found prior beliefs to be more important than language use and argument quality.

Also focusing on persuasion strategies, Wang *et al.* (2019) collect a dataset with 1,017 dialogues, of which 300 have been annotated[v] regarding 10 possible strategies, divided into *persuasive appeal* and *persuasive inquiry*. They then analyze both persuaders' and persuadees' dialogue acts and check how strategies evolve during dialogue turns. A significant number of dialogue acts (40%) have been found to be non-strategic.

Ghosh *et al.* (2014) have carried out an annotation project of blog postings, in which they employed a simplistic *callouts and targets* model, based on a so-called *Pragmatic Argumentation Theory*. This theory states that argument arises from calling out some aspect of a prior contribution. As such, a callout occurs when a subsequent post addresses (part of) a prior post (the target), responding to it. The task of a set of expert annotators was to find each instance of a callout, determine its boundaries and link it to the most recent target, whose boundaries must also be determined. In a follow-up annotation task, crowdsourced annotators have been asked to label the (dis)agreement relation between a callout and its corresponding target and to identify the stance and rationale in callouts identified by expert annotators.

Stab *et al.* (2018) collected a corpus[w] from web-searched documents on eight controversial topics. The top 50 results in each topic have been segmented into sentences, and crowdsourced annotators have been asked to label each sentence as *supporting argument*, *opposing argument*, or *non-argument*, in a flat annotation model. In the corpus, the number of non-arguments largely exceeds the number of supporting or opposing arguments, which is a consequence of the fact that only arguments consisting of individual sentences are taken into account. Several other corpora for stance detection are available, namely those listed in Hardalov *et al.* (2021).

[s] https://nlds.soe.ucsc.edu/iac2.

[t] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2427.

[u] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2428.

[v] https://gitlab.com/ucdavisnlp/persuasionforgood.

[w] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2345.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Park and Cardie (2014, 2018) analyze propositions in online user comments regarding governmental rules and policies. With the aim of assessing how adequately the arguments conveyed are supported, propositional units are classified as one of *non-experiential fact*, *experiential fact* (testimony), *value*, *policy*, or *reference to a resource*. Support relations between propositions include *reason* (if a proposition provides a rationale for another) and *evidence* (when a proposition proves whether another proposition is true or not). The resulting Cornell eRulemaking Corpus[x] consists of 731 comments, 4931 elementary units, and 1221 support relations.

Goudas *et al.* (2014) collect 204 Greek social media documents concerning renewable energy sources and manually annotate them with domain entities and text segments corresponding to argument premises. In their approach, claims are implied by the identified domain entities about which an author is arguing by expressing positive or negative views.

The scientific articles text genre has also been explored in a few projects. Kirschner, Eckle-Kohler, and Gurevych (2015) analyzed 24 German-written articles from educational research. Focusing on the introduction and discussion sections of such articles, they followed a sentence-level annotation scheme, considering four types of relations: two argumentative (*support* and *attack*) and two discursive (*detail* and *sequence*). Lauscher, Glavaš, and Ponzetto (2018b), conversely, have explored a fine-grained model for annotating argumentative components and their relations when addressing a corpus of 40 English publications on computer graphics. Although the authors have considered employing the Toulmin model of argumentation, they have found that most of its elements are not present in this text genre and observe that relations between argumentative components can be of a different nature. They distinguish two types of claims (*own* and *background*), together with a *data* component (a fact that serves as evidence for or against a claim). They also consider three types of relations: *support*, *contradict*, and *semantically same* (a kind of argument coreference). In their approach, argumentative components can be of arbitrary lengths, from a single token to multiple sentences.

Looking at peer-reviews of scientific articles, Hua *et al.* (2019) segment a corpus on propositions that are classified in one of *evaluation*, *request*, *fact*, *reference*, *quote*, or *non-arg*. Their AMPERE corpus is composed of 400 reviews with 10,386 propositions.

Several other annotated corpora exist with different kinds of argument-related annotations. We refer the reader to the surveys published in the past (Lippi and Torroni 2016; Cabrio and Villata 2018; Lawrence and Reed 2019; Schaefer and Stede 2021). Other useful sources of corpora for several argument mining-related tasks include IBM's Project Debater data sets repository, TU Darmstadt's argument mining corpora repository, Webis argumentation data repository, or the Natural Language and Dialogue Systems corpora website.

4. On corpora compatibility

The evident variety of theoretical groundings of existing argument-annotated corpora, as seen in Section 3.2, hinders their joint usage in approaches that benefit from a large annotated corpus, as is the case of modern deep learning techniques applied in argument mining tasks. The fact that many other annotation efforts follow ad-hoc specifications of argument components or concepts, as seen in Section 3.3, worsens this problem. Some attempts to make argumentation models and corpora interoperable in some sense have been pursued, which we revisit in this section. We also discuss the challenges that lie ahead when trying to harmonize corpora that also differ in the underlying textual genres (such as those discussed in Section 2.1).

4.1. Format

One can think of interoperable or compatible argument corpora at two distinct levels. One of them is the *syntactical* level, related to how arguments are actually represented in a language or

[x] https://facultystaff.richmond.edu/~jpark/data/jpark_lrec18.zip.

formalism. Given the broad range of disciplines in which argumentation has been studied over time (including philosophy, linguistics, artificial intelligence, and multi-agent systems), it is a daunting task to come up with a representation formalism that is able to capture all the particular concerns and biases of different fields of study. Nevertheless, the Argument Interchange Format (AIF) (Chesñevar *et al.* 2006; Rahwan *et al.* 2007; Rahwan and Reed 2009) is a serious effort in that direction and started as a draft specification intended for representation and exchange of data between various argumentation tools and agent-based applications (Chesñevar *et al.* 2006). AIF is a cross-discipline effort to provide an abstract model that can then be extended to specific purposes or domains, enabling the accommodation of several argumentation theories or schemes. AIF aimed at bridging the gap between prior argument markup languages (Reed and Rowe 2004), mostly concerned with visually structuring natural language arguments in a diagrammatic form, and the need to represent formal logic-based argumentation in computational systems, enabling automated argumentation-based reasoning in multi-agent systems.

AIF's core ontology specification includes two types of nodes for defining arguments: information nodes (*I-nodes*) relate to the content of arguments and are thus domain-dependent; scheme nodes (*S-nodes*) involve domain-independent patterns of reasoning, including not only rules of inference in the logical sense but also reasoning steps that are not part of classical logical inference. S-nodes are further specified into inference schemes (*RA-nodes*), preference schemes (*PA-nodes*), and conflict schemes (*CA-nodes*). Edges between nodes can be *scheme edges*, which originate in S-nodes, or *data edges*, which originate in I-nodes and end in S-nodes. I-to-S edges are "data-supplying" edges, while S-to-I are "conclusion" edges; S-to-S edges are "warrant" edges and allow for "meta-reasoning" or "meta-argumentation," for instance when some justification is needed for the application of an inference rule. This core specification is rich enough to encompass any concepts from an extension ontology (including, e.g., notions of support, attack, rebut, or undercut), in the sense that such concepts can, in principle, be described in terms of the core ontology. AIF also touches on issues related to communication (locutions and protocols) and the context of argumentation (its participants, argumentation theory and commitment rules, and so on).

Despite the effort of defining a standard for representing arguments, not all argument annotation projects adopt this formalism, although a repository of corpora using AIF (Lawrence, Janier, and Reed 2015) and software strictly following AIF (Janier *et al.* 2014) exist. Many projects, in particular those relying on a shallower argumentation model, resort to more generic representation formats, namely brat's standoff[y] or a format based on CoNLL (Hajič *et al.* 2009) or BIO encodings (Tjong Kim Sang and Veenstra 1999).

4.2. Content

At the *semantic* level, corpora compatibility is related to the semantics of the underlying argumentation models (which, as we have seen earlier, are sometimes simplified or distorted) and is much trickier than achieving syntactic interoperability. Most formulations of argument schemes can be drilled down to support relations between reasons and claims, and critical questions can be seen as possible attack formulations. At the same time, some theoretical models of argumentation are based on or are modifications of prior existing models. For instance, Freeman's extended standard approach is both a simplification and a refinement of Toulmin's model (see Section 2.2.2). This provides some cues for interchangeability between the two models, although perhaps in a non-bidirectional way. For instance, one can (rather simplistically) see the relations between some of Toulmin's model components as certain kinds of structures in the Freeman model; for example, data, warrant, and claim can be mapped to a linked structure.

[y] https://brat.nlplab.org/standoff.html.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

In fact, interconnections between the two have been explored by Reed and Rowe (2005) in an attempt to achieve theory neutrality in diagramming tools. However, while the "building blocks" or "atoms" in each theory are not that different, Toulmin's framework is semantically deeper than a set of premises and conclusions. Moreover, complex argument structures enabled by the standard approach are not directly translatable to the Toulmin model, which focuses on the six-component individual argument structure. Reed and Rowe suggest that some of these components can be seen as claims in other arguments, allowing for building Toulmin diagrams of arbitrary complexity.

Some theories are considerably distinct in the kinds of elements they intend to capture, and thus, their most distinguishing features cannot be translated back and forth to concepts in other theories. IAT's focus on the anchoring of propositions and reasoning steps in the dialogue seems to capture phenomena that are distinct from those found in other theories. PTA's orthogonal classification of arguments based on propositional content is also somewhat distinct, although we can find some similarities between, on the one hand, arguments of first/second order and the types of propositions, and on the other hand some of Walton's argument schemes.

Visser *et al.* (2021) have matched argument types as obtained using PTA with some of Walton's argument schemes. They developed a co-occurrence matrix on the corpus analyzed that shows some correspondence between, for instance, *1 pre PF* (first-order predicate arguments supporting a proposition of policy based on one of fact) and arguments from *consequences*, *example*, *practical reasoning,* and *pragmatic* argument schemes in Walton's taxonomy.

Translations between argumentation theories will most probably leave out some semantic details that are captured in one theory and not in the others. In some cases, it seems possible to combine features from more than one model to obtain richer annotated corpora. Some approaches have done exactly that by applying an additional annotation layer based on a theory different from the one originally employed for that corpus (Musi *et al.* 2016; Reisert *et al.* 2017b; Musi *et al.* 2018; Visser *et al.* 2021). Work relating argument schemes and discourse relations in Rhetorical Structure Theory (RST) (Mann and Thompson 1988) has also been done (Peldszus and Stede 2013; Musi *et al.* 2018).

One could think that the highest gain would be obtained from merging those corpora into one that could be used to train machine learning models at a larger scale. In fact, most corpora for argument mining in its various tasks are modest in size due to the intricate, demanding, and time-consuming nature of argument annotation, as discussed in Section 3.1. But again, given the variety of annotation guidelines and argumentation granularity that have been pursued in different projects, this is no easy task. Such peculiarities of each argument annotation project are rooted in the textual genre it targets, which then influences the adopted theoretical argumentation model. As a consequence, even if, in principle, one could be able to harmonize the differences in argument annotation of the different projects, we would get a heterogeneous corpus encompassing documents that differ significantly in argumentative style—from those strongly marked with discourse markers and with a well-defined structure (such as legal texts or argumentative essays) to those not assuming too much about writing rules (such as opinion articles and social media content).

Bringing together such variety in a joint corpus naively assumes that we can tame argumentation with a one-size-fits-all approach. Moreover, such a heterogeneous corpus would likely hinder rather than help the potential of machine learning to address the supervised training of fine-grained argument mining models successfully. There is evidence that recent large language models can exploit varied corpora for several kinds of tasks; at the same time, some work exists that jointly makes use of cross-domain data for shallower argumentation tasks, such as stance detection (Hardalov *et al.* 2021). However, we argue that properly addressing fine-grained argument mining in a cross-genre fashion is much more involved—evidence of which is the existence of such diverse argumentation models, such as those portrayed in Section 2.2.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

4.3. Practical compatibility

The usefulness of different annotation schemes for disparate albeit related tasks should also be considered in light of recent advancements in natural language processing and machine learning. It is now a standard approach to rely on pre-trained models and fine-tune them using labeled data for the target task (Devlin *et al.* 2019; Radford *et al.* 2019; Yang *et al.* 2019; Conneau *et al.* 2020). This ranges from pre-trained language models (Devlin *et al.* 2019; Radford *et al.* 2019; Brown *et al.* 2020) to multi-task (Liu *et al.* 2017; Augenstein, Ruder, and Søgaard 2018; (Radford *et al.* 2019; Ruder *et al.* 2019a; Pfeiffer *et al.* 2020) and transfer learning between tasks (Ruder 2017; Schröder and Biemann 2020; Sharma, Zheng, and Awadallah 2021). Cross-lingual approaches are also widespread, using techniques such as multilingual word embeddings (Chen and Cardie 2018; Ruder, Vulić, and Søgaard 2019b), language models (Devlin *et al.* 2019; Conneau *et al.* 2020), and transfer learning (Eger *et al.* 2018; Schuster *et al.* 2019).

The main assumption in all these approaches is that we can start from a prior model that has been trained either in an unsupervised way with large amounts of data or with labeled data from an auxiliary task, taken to be related to the task that we want the final model to address—the target task. Typically, the amount of data available from that auxiliary task is much higher than what we can afford to have in our task of interest. The model is then retrained (or in some cases simultaneously trained) with the (usually much less) available data for the desired target task.

A critical issue is how to identify the auxiliary tasks that are most useful for the job, which amounts to having some notion of similarity between tasks (Ruder 2017; Schröder and Biemann 2020). Multi-task learning usually consists of learning representations that are useful across tasks, often achieved through hard parameter sharing of hidden layers in neural networks (Augenstein *et al.* 2018). Some approaches explore splits between shared and private spaces in such representations (Liu *et al.* 2017), with the aim of refining the shared space and thus making it more useful for the target task. The similarity between the chosen tasks is a determinant of which approach is more promising—and also whether a multi-task or pure transfer-learning approach is a better choice (Pruksachatkun *et al.* 2020; Poth *et al.* 2021). Usually, the closer the tasks are (in terms of semantics and label spaces), the better the chances that joint multi-task learning is viable.

When employing multi-task or transfer learning in argument-annotated corpora, one must bear in mind that besides facing differences in the kinds of annotations that have been produced, we are usually in the presence of a multi-faceted domain shift. While domain adaptation for NLP tasks has been addressed in the past (Blitzer, Dredze, and Pereira 2007; Ramponi and Plank 2020), in the case of argument-annotated corpora a change in domain may imply dealing with different topics, styles, genres, or linguistic register (Plank 2016; Ramponi and Plank 2020), which may reflect very different argumentation styles and strategies (namely the presence or absence of argumentative discourse markers), as discussed in Section 2.1.

Addressing the problem of genre shift in different corpora may be approached using text-to-text style conversion techniques (Fu *et al.* 2018; Raffel *et al.* 2020; Hu, Lee, and Aggarwal 2020), where assuring semantic content preservation when converting between styles or genres is a critical issue. However, many of these approaches have been applied to tasks such as polarity inversion in reviews or informal to formal language conversion. At the same time, work on natural language generation of argumentative text is a very active area of research (Elhadad 1992; Zukerman, McConachy, and George 2000; Reisert *et al.* 2015; Hua and Wang 2018; Cerutti, Toniolo, and Norman 2019; Alshomary *et al.* 2021; Schiller, Daxenberger, and Gurevych 2021a).

Discourse markers are particularly useful for argument mining, although doing so often produces models that are not robust and can be easily tricked (Opitz and Frank 2019). That is because the reasoning steps underlying an argumentative text are supposed to rely much more on the content of the argumentative discourse units and how they relate to each other than on the usage of strong argumentative markers—which may be more or less prevalent, depending on the text genre. A direct way of harmonizing different corpora in this regard is thus to strip out discourse markers altogether or to inject them wherever appropriate (Elhadad 1992; Cerutti, Toniolo, and Norman 2019).

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Other loosely related corpora (i.e., those which have not been annotated with argumentation structures but with other linguistic phenomena) can be used in auxiliary tasks. Examples include discourse relations in the Penn Discourse TreeBank (Prasad *et al.* 2008; Hewett *et al.* 2019) (such as subordinating and coordinating conjunctions, adverbials, and several kinds of implicit relations) and textual entailment/natural language inference (Cabrio and Villata 2012, 2013; Choi and Lee 2018; Conneau *et al.* 2018).

5. Discussion

Establishing a connection between theoretical argumentation models and actual argumentation practice is not an easy or straightforward task. Even though some argumentation models follow a bottom-up approach, obtaining large corpora with rich argument-related annotations is notoriously hard. For that reason, most existing corpora are relatively small in size.

We can draw some useful insights from analyzing existing argument annotation outputs. One is that the more intricate and semantically demanding the theories are, the harder it will be to bring about quality annotated corpora of significant sizes. We include Toulmin's, AMT, IAT, and, to some extent, Walton's argumentation schemes in this group. The latter usually requires a careful selection of schemes and the development of rigorous guidelines to help the annotator find matching instances. Generally, the less consensual the argumentation model (see, e.g., Freeman's critiques on Toulmin's model, or the different groupings of Walton's schemes), the less agreement is likely to be observed as an outcome of the annotation process. This is due to the model being harder to instantiate or amenable to different interpretations.

Freeman's focus on the external structure of arguments makes it a semantically less detailed model and one that is easier to employ when compared to argumentation schemes. Furthermore, one can use parts of the model without significantly harming its nature. For instance, several annotation projects described in Section 3.2.2 have avoided certain parts of the model, such as types of structures, modalities, or defeaters; still, the essential elements of conclusions, premises, and relations among them remain—the macrostructure. PTA is also easier to employ due to its low complexity and orthogonal nature—the three layers of this model (predicate/subject propositional content, first-order/second-order arguments, and propositions of policy/value/fact) are easy to understand and identify in concrete arguments. It also has the benefit that each layer (i.e., each partial annotation) is valuable in itself.

Text genre also plays a significant role in the (manual) mining of arguments. Texts denoting a less explicit and more subtle argumentation style, relying heavily on enthymemes, or requiring more refined world knowledge (Lauscher *et al.* 2022) are significantly harder to analyze. Argumentation models aside, the most challenging task is telling apart argumentative excerpts from merely descriptive content, as argument density tends to be lower than in other genres with explicit argumentative discourse markers.

To further ease the task of mining arguments, some annotation projects have produced their own corpora in a controlled text generation fashion or have used student-written essays on controversial topics, making it easier to annotate. We argue that these kinds of approaches will inevitably limit the ability of computational models trained on these corpora to generalize to real-world data, which is seldom as clean. An interesting line of research consists of using the argumentation models as a starting point to automatically produce natural language arguments from structured data. Combined with conditioned text generation based on style information, this generation process may be very promising in building self-annotated argument corpora of diverse genres that are big enough to train computational models on.

The granularity of argument annotations in existing corpora—both those explicitly following a theoretical argumentation model (identified in Section 3.2) and those that do not (such as the ones briefly described in Section 3.3)—conditions the kinds of applications one can envision

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

when making use of these valuable resources. In fact, promising applications of the study of argumentation in practice are manifold. Argument mining has been presented as a natural extension to sentiment analysis and opinion mining (Grosse *et al.* 2015; Dragoni *et al.* 2018; Lytos *et al.* 2019; Chen *et al.* 2021). Understanding people's arguments concerning, a given topic is important from several perspectives, such as determining stance (Allaway and McKeown 2020; Küçük and Can 2020; Schiller, Daxenberger, and Gurevych 2021b), argument strength (Habernal and Gurevych 2016a; Wachsmuth *et al.* 2017), and the presence of fallacies (Visser *et al.* 2017) or semantic incongruity.

From a forensic linguistics perspective, we can envision employing techniques such as rhetorical profiling (Visser *et al.* 2021), which aims at characterizing—or profiling from a sociolinguistic point of view—speakers or authors in terms of arguing style, including their preference for certain types of argument schemes. In a similar vein, by exploring argumentation trails, one can develop refined models for detecting rhetorical strategies employed in fora such as news diffusion channels (Al-Khatib *et al.* 2016), or prejudicial bias (Spliethöver and Wachsmuth 2020) in legal texts (Pinto *et al.* 2020).

With the aim of advancing computational approaches to understand text-based argumentation, some shared tasks have been proposed, such as the Argument Reasoning Comprehension Task (Habernal *et al.* 2018), which focuses on identifying implicit warrants. Given the focus of state-of-the-art systems on deep learning approaches (and in particular Transformer-based architectures), it has been shown that current models are unable to capture a deep understanding of how argumentation unfolds, easily getting trapped in data spuriousness that allows such models to perform well in terms of the adopted evaluation metrics and on the provided test sets (Branco *et al.* 2021). This phenomenon is known as shortcut learning (Geirhos *et al.* 2020). Key point analysis and summarization (Bar-Haim *et al.* 2020a,b) is another recently proposed argument-related shared task: given a corpus of texts focusing on a topic of interest, the aim is to extract the most relevant key-points, together with their relative prevalence. The proposal of new tasks that harness the richness of different argumentation models may be instrumental in fostering research in this domain.

With a focus on argument retrieval, the Touché shared tasks have also been proposed (Bondarenko *et al.* 2021, 2022). However, most participating models address the problem by employing information retrieval techniques rather than focusing on argumentation models.

6. Conclusions

Each argument annotation project has its own aim, typically related to the tasks (argument mining or otherwise) that can be performed with the resulting corpora. The choice of argumentation model to employ should be, at least in part, determined by that aim—such choice will highly constrain the (argument mining) tasks that the corpus will support. The diversity of existing argument annotation projects is directly linked to the range of argumentation models that have been followed. Additionally, the chosen argumentation model has a significant impact on the usefulness of the obtained corpus, measured both in terms of its qualitative nature and of the inter-annotation agreement. In fact, more sophisticated argumentation models, while being more informative (as they allow capturing richer argumentative functions), are also more challenging for annotators, who must go through very specific annotation guidelines. This usually requires intensive training. Furthermore, the adoption of different argumentation models is an obstacle both for corpora comparison and for exploiting their possible complementarity.

Differences aside, fine-grained corpus-based argument mining is still a long way from achieving useful results, perhaps except for formal and highly structured argumentative texts (such as in the legal Mochales and Moens 2011 or academic Lauscher *et al.* 2018a domains, including essays Stab and Gurevych 2014; Peldszus and Stede 2016). The usefulness of argument annotation

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

is not limited to argument mining, though. Argument diagrams (Reed, Walton, and Macagno 2007) have long been used in education (van Gelder 2001; Kirschner *et al.* 2003), deliberation (Karamanou *et al.* 2011) and critical thinking (Green, Branon, and Roosje 2019). Given the increasing attention that these and closely related topics are getting (such as fact-checking and forensic linguistics), a wider range of applications is likely to come to light in the near future.

Acknowledgments. The first author would like to thank the Centre for Argument Technology (ARG-tech), in particular Chris Reed and Jacky Visser, for hosting a visit back in 2019, based on which this work has come to light. The authors would like to thank the anonymous reviewers whose comments have resulted in significant improvements of the original manuscript. This research is supported by LIACC (FCT/UID/CEC/0027/2020), CLUP (UIDB/00022/2020) and INESC-ID (UIDB/50021/2020), and is part of the DARGMINTS project (POCI/01/0145/FEDER/031460), funded by Fundação para a Ciência e a Tecnologia (FCT).

Conflicts of interest. The authors declare none.

References

Abbott R., Ecker B., Anand P. and Walker M. (2016). Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA), pp. 4445–4452.

Aharoni E., Polnarov A., Lavee T., Hershcovich D., Levy R., Rinott R., Gutfreund D. and Slonim N. (2014). A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland. Association for Computational Linguistics, pp. 64–68.

Aharonov R. and Slonim N. (2019). Watch IBM's AI System Debate a Human Champion Live at Think 2019. https://www.ibm.com/blogs/research/2019/02/ai-debate-think-2019/ (accessed 14 April 2021).

Al-Khatib K., Hou Y. and Stede M. (2021). *Proceedings of the 8th Workshop on Argument Mining*, Punta Cana, Dominican. Association for Computational Linguistics Republic.

Al-Khatib K., Wachsmuth H., Kiesel J., Hagen M. and Stein B. (2016). A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan. The COLING 2016 Organizing Committee, pp. 3433–3443.

Allaway E. and McKeown K. (2020). Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 8913–8931.

Alshomary M., Chen W.-F., Gurcke T. and Wachsmuth H. (2021). Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics, pp. 224–233.

Augenstein I., Ruder S. and Søgaard A. (2018). Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume* 1 *(Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1896–1906.

Bal B.K. and Saint Dizier P. (2010). Towards building annotated resources for analyzing opinions and argumentation in news editorials. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA), pp. 1152–1158.

Bar-Haim R., Eden L., Friedman R., Kantor Y., Lahav D. and Slonim N. (2020a). From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 4029–4039.

Bar-Haim R., Kantor Y., Eden L., Friedman R., Lahav D. and Slonim N. (2020b). Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 39–49.

Basile V., Fell M., Fornaciari T., Hovy D., Paun S., Plank B., Poesio M. and Uma A. (2021). We need to consider disagreement in evaluation. In *Procs 1st Workshop on Benchmarking: Past, Present and Future,* Online. ACL, pp. 15–21.

Bayerl P.S. and Paul K.I. (2011). What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics* **37**(4), 699–725.

Bentahar J., Moulin B. and Bélanger M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* **33**(3), 211–259.

Besnard P. and Hunter A. (2001). A logic-based theory of deductive arguments. *Artificial Intelligence* **128**(1–2), 203–235.

Besnard P. and Hunter A. (2008). *Elements of Argumentation*. Cambridge, MA: The MIT Press.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Bhatia V.K. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Routledge.

Biber D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Blair J.A. (2003). Relationships among logic, dialectic and rhetoric. In *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*. Dordrecht: Springer Netherlands, pp. 91–107.

Blitzer J., Dredze M. and Pereira F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. Association for Computational Linguistics, pp. 440–447.

Bondarenko A., Fröbe M., Kiesel J., Syed S., Gurcke T., Beloucif M., Panchenko A., Biemann C., Stein B., Wachsmuth H., Potthast M. and Hagen M. (2022). Overview of touché 2022: Argument retrieval: Extended abstract. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, Berlin, Heidelberg: Springer-Verlag, pp. 339–346.

Bondarenko A., Gienapp L., Fröbe M., Beloucif M., Ajjour Y., Panchenko A., Biemann C., Stein B., Wachsmuth H., Potthast M. and Hagen M. (2021). Overview of touché 2021: Argument retrieval: Extended abstract. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, pp. 574–582.

Branco R., Branco A., António Rodrigues J. and Silva J.R. (2021). Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 1504–1521.

Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I. and Amodei D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F. and Lin H. (eds), *Advances in Neural Information Processing Systems*, vol. 33, Online. Curran Associates, Inc., pp. 1877–1901.

Budzynska K., Janier M., Reed C., Saint-Dizier P., Stede M. and Yakorska O. (2014). A model for processing illocutionary structures and argumentation in debates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 917–924.

Budzynska K. and Reed C. (2011). Whence inference. Technical report, University of Dundee.

Budzynska K. and Villata S. (2018). Processing natural language argumentation. In Baroni P., Gabbay D., Giacomin M. and van der Torre L. (eds), *Handbook of Formal Argumentation*. Milton Keynes, UK: College Publications, pp. 577–627.

Burstein J., Kukich K., Wolff S., Lu C. and Chodorow M. (1998). Enriching automated essay scoring using discourse marking. In *Discourse Relations and Discourse Markers (Proceedings of the Workshop)*, Montreal, Quebec, Canada, pp. 15–21.

Cabrio E. and Villata S. (2012). Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea. Association for Computational Linguistics, pp. 208–212.

Cabrio E. and Villata S. (2013). A natural language bipolar argumentation approach to support users in online debate interactions†. *Argument & Computation* **4**(3), 209–230.

Cabrio E. and Villata S. (2018). Five years of argument mining: A data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization, pp. 5427–5433.

Cerutti F., Toniolo A. and Norman T. (2019). On natural language generation of formal argumentation. In Santini, F. and Toniolo, A. (eds), *Proceedings of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2019)*, CEUR Workshop Proceedings, Rende, Italy. Sun SITE Central Europe, pp. 15–29.

Chen C.-C., Huang H.-H. and Chen H.-H. (2021). *From Opinion Mining to Financial Argument Mining*. *SpringerBriefs in Computer Science*. Singapore: Springer.

Chen X. and Cardie C. (2018). Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 261–270.

Chesñevar C., McGinnis J., Modgil S., Rahwan I., Reed C., Simari G., South M., Vreeswijk G. and Willmott S. (2006). Towards an argument interchange format. *Knowledge Engineering Review* **21**(4), 293–316.

Choi H. and Lee H. (2018). GIST at SemEval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 773–777.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 8440–8451.

Conneau A., Rinott R., Lample G., Williams A., Bowman S., Schwenk H. and Stoyanov V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 2475–2485.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Das D. and Taboada M. (2013). Explicit and implicit coherence relations: A corpus study. In *Proceedings of the 2013 Annual Conference of the Canadian Linguistic Association*.

Das D. and Taboada M. (2018). Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes* **55**(8), 743–770.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume* 1 *(Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.

do Carmo R. (2012). *Linguagem, Argumentação e Decisão Judiciária*. Coimbra: Coimbra Editora.

Dragoni M., Da Costa Pereira C., Tettamanzi A. G. B. and Villata S. (2018). Combining argumentation and aspect-based opinion mining: The SMACk system. *AI Communications* **31**(1), 75–95.

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

Durmus E. and Cardie C. (2018). Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume* 1 *(Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1035–1045.

Duthie R., Budzynska K. and Reed C. (2016). Mining ethos in political debate. In Baroni P., Gordon T., Scheffler T. and Stede M. (eds), *Computational Models of Argument*. Frontiers in Artificial Intelligence and Applications, vol. 287, Netherlands. IOS Press, pp. 299–310.

Egawa R., Morio G. and Fujita K. (2019). Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy. Association for Computational Linguistics, pp. 422–428.

Eger S., Daxenberger J., Stab C. and Gurevych I. (2018). Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 831–844.

Elhadad M. (1992). Generating coherent argumentative paragraphs. In *COLING 1992 Volume* 2: *The 14th International Conference on Computational Linguistics*, Nantes, France.

Eliot L. (2021). Identifying a Set of Autonomous Levels for AI-Based Computational Legal Reasoning. *MIT Computational Law Report*. https://law.mit.edu/pub/identifyingasetofautonomouslevelsforaibasedcomputationallegalreasoning.

Freeman J.B. (1991). *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. Pragmatics and Discourse Analysis Series. Berlin and New York: Foris Publications.

Freeman J.B. (2011). *Argument Structure: Representation and Theory*. Argumentation Library. Dordrecht: Springer Netherlands.

Fu Z., Tan X., Peng N., Zhao D. and Yan R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, vol. 32, New Orleans.

Geirhos R., Jacobsen J.-H., Michaelis C., Zemel R., Brendel W., Bethge M. and Wichmann F.A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673.

Ghosh D., Muresan S., Wacholder N., Aakhus M. and Mitsui M. (2014). Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland. Association for Computational Linguistics, pp. 39–48.

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

Goudas T., Louizos C., Petasis G. and Karkaletsis V. (2014). Argument extraction from news, blogs, and social media. In Likas A., Blekas K. and Kalles D. (eds), *Artificial Intelligence: Methods and Applications*. Cham: Springer International Publishing, pp. 287–299.

Govier T. (2010). *A Practical Study of Argument*, 7th Edn. Boston: Cengage Learning.

Green N.L., Branon M. and Roosje L. (2019). Argument schemes and visualization software for critical thinking about international politics. *Argument & Computation* **10**(1), 41–53.

Grosse K., González M., Chesñevar C. and Maguitman A. (2015). Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Communications* **28**, 387–401.

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*, Austin, Texas. Association for Computational Linguistics, pp. 1214–1223.

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)*, Berlin, Germany. Association for Computational Linguistics, pp. 1589–1599.

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

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Habernal I., Wachsmuth H., Gurevych I. and Stein B. (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume* 1 *(Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1930–1940.

Haddadan S., Cabrio E. and Villata S. (2018). Annotation of argument components in political debates data. In Sandra Kübler, H.Z. (ed.), *Proceedings of the Workshop on Annotation in Digital Humanities*, Sofia, Bulgaria. CEUR Workshop Proceedings, pp. 12–16.

Haddadan S., Cabrio E. and Villata S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 4684–4690.

Hajič J., Ciaramita M., Johansson R., Kawahara D., Mart M.A., Màrquez L., Meyers A., Nivre J., Padó, S., Štěpánek J., Straňák P., Surdeanu M., Xue N. and Zhang Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, Boulder, Colorado. Association for Computational Linguistics, pp. 1–18.

Hansen H. and Walton D. (2013). Argument kinds and argument roles in the ontario provincial election, 2011. *Journal of Argumentation in Context* **2**, 226–258.

Hardalov M., Arora A., Nakov P. and Augenstein I. (2021). Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 9011–9028.

Hargie O., Dickson D. and Tourish D. (2004). *Communication Skills for Effective Management*. Basingstoke: Palgrave.

Hautli-Janisz A., Kikteva Z., Siskou W., Gorska K., Becker R. and Reed C. (2022). Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Hewett F., Prakash Rane R., Harlacher N. and Stede M. (2019). The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy. Association for Computational Linguistics, pp. 98–103.

Hidey C., Musi E., Hwang A., Muresan S. and McKeown K. (2017). Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 11–21.

Hu Z., Lee R.K. and Aggarwal C.C. (2020). Text style transfer: A review and experiment evaluation. CoRR, abs/2010.12742.

Hua X., Nikolov M., Badugu N. and Wang L. (2019). Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 2131–2137.

Hua X. and Wang L. (2018). Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume* 1*: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 219–230.

Huber R. and Snider A.C. (2006). *Influencing Through Argument*, updated Edn. New York: International Debate Education Association.

Janier M., Lawrence J. and Reed C. (2014). OVA+: An argument analysis interface. In Parsons S., Oren N., Reed C. and Cerutti F. (eds), *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, Pitlochry. IOS Press, pp. 463–464.

Janier M. and Reed C. (2017). Towards a theory of close analysis for dispute mediation discourse. *Argumentation* **31**(1), 45–82.

Karamanou A., Loutas N. and Tarabanis K. (2011). Argvis: Structuring political deliberations using innovative visualisation technologies. In Tambouris E., Macintosh A. and de Bruijn H. (eds), *Electronic Participation*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 87–98.

Katzav J. and Reed C.A. (2004). On argumentation schemes and the natural classification of arguments. *Argumentation* **18**(2), 239–259.

Kerbrat-Orecchioni C. (2004). Introducing polylogue. *Journal of Pragmatics* **36**(1), 1–24.

Kienpointner M. (1986). Towards a typology of argument schemes. In *Argumentation: Across the Lines of Discipline, Proceedings of the Conference on Argumentation*. Amsterdam: Amsterdam University Press, pp. 275–287.

Kienpointner M. (1992). How to classify arguments. In *Argumentation Illuminated*. Amsterdam: Amsterdam University Press, pp. 178–188.

Kirschner C., Eckle-Kohler J. and Gurevych I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO. Association for Computational Linguistics, pp. 1–11.

Kirschner P.A., Shum S.J.B. and Carr C.S. (eds) (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. London, UK: Springer.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Küçük D. and Can F. (2020). Stance detection: A survey. *ACM Computing Surveys* **53**(1), 1–37.

Lauscher A., Glavaš G. and Eckert K. (2018a). ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, Brussels, Belgium. Association for Computational Linguistics, pp. 22–28.

Lauscher A., Glavaš G. and Ponzetto S.P. (2018b). An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, Brussels, Belgium. Association for Computational Linguistics, pp. 40–46.

Lauscher A., Wachsmuth H., Gurevych I. and Glavaš G. (2022). Scientia Potentia Est–On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics* **10**, 1392–1422.

Lawrence J., Janier M. and Reed C. (2015). Working with open argument corpora. In *Proceedings of the 1st European Conference on Argumentation (ECA 2015)*, Lisbon. College Publications.

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

Lewiński M. and Aakhus M. (2014). Argumentative polylogues in a dialectical framework: A methodological inquiry. *Argumentation* **28**(2), 161–185.

Lewiński M. and Mohammed D. (2019). The 2015 paris climate conference: Arguing for the fragile consensus in global multilateral diplomacy. *Journal of Argumentation in Context* **8**(1), 65–90.

Lippi M. and Torroni P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology* **16**(2), 10:1–10:25.

Liu P., Qiu X. and Huang X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume* 1: *Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 1–10.

Lumer C. (2011). Argument schemes – an epistemological approach. In *Argumentation: Cognition and Community. Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pp. 1–32.

Lumley T. and McNamara T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing* **12**(1), 54–71.

Lytos A., Lagkas T., Sarigiannidis P. and Bontcheva K. (2019). The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* **56**(6), 102055.

Macagno F., Walton D. and Reed C. (2017). Argumentation schemes: History, classifications, and computational applications. *Journal of Logics and their Applications* **4**(8), 2493–2556.

Mann W.C. and Thompson S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text* **8**(3), 243–281.

Menini S., Cabrio E., Tonelli S. and Villata S. (2018). Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 4889–4896.

Menini S., Nanni F., Ponzetto S.P. and Tonelli S. (2017). Topic-based agreement and disagreement in US electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 2938–2944.

Mochales R. and Ieven A. (2009). Creating an argumentation corpus: Do theories apply to real arguments?: A case study on the legal argumentation of the ECHR. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL'09, New York, NY, USA. ACM, pp. 21–30.

Mochales R. and Moens M. (2011). Argumentation mining. *Artificial Intelligence and Law* **19**(1), 1–22.

Moens M.-F. (2018). Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument and Computation* **9**, 1–14.

Musi E. and Aakhus M. (2018). Discovering argumentative patterns in energy polylogues: A macroscope for argument mining. *Argumentation* **32**(3), 397–430.

Musi E., Ghosh D. and Muresan S. (2016). Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, Berlin, Germany. Association for Computational Linguistics, pp. 82–93.

Musi E., Stede M., Kriese L., Muresan S. and Rocci A. (2018). A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Nguyen Q.V., Duong C.T., Nguyen T.T., Weidlich M., Aberer K., Yin H. and Zhou X. (2017). Argument discovery via crowdsourcing. *The VLDB Journal* **26**(4), 511–535.

Nute D. (1994). Defeasible logic. In Gabbay D.M., Hogger C.J. and Robinson J.A. (eds), *Handbook of Logic in Artificial Intelligence and Logic Programming (Vol.* 3). New York, NY, USA: Oxford University Press, Inc., pp. 353–395.

O'Keefe D.J. (1977). Two concepts of argument. *The Journal of the American Forensic Association* **13**(3), 121–128.

O'Neill J.M., Laycock C. and Scales R.L. (1925). *Argumentation and Debate*. London: MacMillan.

Opitz J. and Frank A. (2019). Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy. Association for Computational Linguistics, pp. 25–34.

Park J. and Cardie C. (2014). Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, Baltimore, Maryland. Association for Computational Linguistics, pp. 29–38.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Park J. and Cardie C. (2018). A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Peldszus A. and Stede M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence* **7**(1), 1–31.

Peldszus A. and Stede M. (2016). An annotated corpus of argumentative microtexts. In Mohammed D. and Lewinski M. (eds), *Argumentation and Reasoned Action - Proceedings of the 1st European Conference on Argumentation, Lisbon, 2015*. London: College Publications.

Perelman C. and Olbrechts-Tyteca L. (1958). *La nouvelle rhétorique. Traité de l'argumentation*. Paris: Presses Universitaires de France.

Perelman C. and Olbrechts-Tyteca L. (1969). *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, IN: University of Notre Dame Press.

Pfeiffer J., Vulić I., Gurevych I. and Ruder S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 7654–7673.

Pinto A.G., Lopes Cardoso H., Duarte I.M., Warrot C.V. and Sousa-Silva R. (2020). Biased language detection in court decisions. In Analide C., Novais P., Camacho D. and Yin H. (eds), *Intelligent Data Engineering and Automated Learning – IDEAL 2020*. Cham: Springer International Publishing, pp. 402–410.

Plank B. (2016). What to do about non-standard (or non-canonical) language in NLP. In Dipper S., Neubarth F. and Zinsmeister H. (eds), *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016*. Bochumer Linguistische Arbeitsberichte, vol. 16, Bochum, Germany.

Pollock J.L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA, USA: MIT Press.

Poth C., Pfeiffer J., Rücklé A. and Gurevych I. (2021). What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 10585–10605.

Poudyal P., Savelka J., Ieven A., Moens M.F., Goncalves T. and Quaresma P. (2020). ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, Online. Association for Computational Linguistics, pp. 67–75.

Prasad R., Dinesh N., Lee A., Miltsakaki E., Robaldo L., Joshi A. and Webber B. (2008). The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Pruksachatkun Y., Phang J., Liu H., Htut P.M., Zhang X., Pang R.Y., Vania C., Kann K. and Bowman S.R. (2020). Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 5231–5247.

Pustejovsky J. and Stubbs A. (2012). *Natural Language Annotation for Machine Learning*. Sebastopol, CA: O'Reilly.

Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I. (2019). Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W. and Liu P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67.

Rahwan I., Ramchurn S.D., Jennings N.R., Mcburney P., Parsons S. and Sonenberg L. (2003). Argumentation-based negotiation. *Knowledge Engineering Review* **18**(4), 343–375.

Rahwan I. and Reed C. (2009). The argument interchange format. In *Argumentation in Artificial Intelligence*. Dordrecht: Springer, pp. 383–402.

Rahwan I., Zablith F. and Reed C. (2007). Laying the foundations for a world wide argument web. *Artificial Intelligence* **171**(10), 897–921.

Ramponi A. and Plank B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics, pp. 6838–6855.

Reed C. and Budzynska K. (2011). How dialogues create arguments. In *Proceedings of the 7th Conference of the International Society for the Study of Argumentation.*

Reed C. and Long D. (1997). Persuasive monologue. In *Proceedings of the 2nd International Conference of the Ontario Society for the Study of Argumentation (OSSA).*

Reed C., Palau R.M., Rowe G. and Moens M.-F. (2008). Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Reed C. and Rowe G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools* **14**, 961–980.

Reed C. and Rowe G. (2005). Translating toulmin diagrams: Theory neutrality in argument representation. *Argumentation* **19**(3), 267–286.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Reed C. and Walton D. (2003). Argumentation schemes in argument-as-process and argument-as-product. In *Proceedings of the Conference Celebrating Informal Logic @25*, Windsor, ON.

Reed C., Walton D. and Macagno F. (2007). Argument diagramming in logic, law and artificial intelligence. *Knowledge Engineering Review* **22**(1), 87–109.

Reisert P., Inoue N., Okazaki N. and Inui K. (2015). A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Denver, CO. Association for Computational Linguistics, pp. 45–55.

Reisert P., Inoue N., Okazaki N. and Inui K. (2017a). A corpus of deep argumentative structures as an explanation to argumentative relations. arXiv:1712.02480.

Reisert P., Inoue N., Okazaki N. and Inui K. (2017b). Deep argumentative structure analysis as an explanation to argumentative relations. In *Proceedings of The 23rd Annual Meeting of the Association for Natural Language Processing*, pp. 38–41.

Rigotti E. and Greco S. (2019). *Inference in Argumentation: A Topics-Based Approach to Argument Schemes*. Argumentation Library. Cham: Springer.

Rigotti E. and Greco Morasso S. (2009). Argumentation as an object of interest and as a social and cultural resource. In Muller Mirza N. and Perret-Clermont A.-N. (eds), *Argumentation and Education: Theoretical Foundations and Practices*. Boston, MA: Springer US, pp. 9–66.

Rigotti E. and Greco Morasso S. (2010). Comparing the argumentum model of topics to other contemporary approaches to argument schemes: The procedural and material components. *Argumentation* **24**(4), 489–512.

Rissland E. (1988). Artificial intelligence and legal reasoning: A discussion of the field and gardner's book. *AI Magazine* **9**(3), 45.

Rocha G., Leite B., Trigo L., Cardoso H.L., Sousa-Silva R., Carvalho P., Martins B. and Won M. (2022a). Predicting argument density from multiple annotations. In Rosso P., Basile V., Martínez R., Métais E. and Meziane F. (eds), *Natural Language Processing and Information Systems*. Cham: Springer International Publishing, pp. 227–239.

Rocha G., Trigo L., Lopes Cardoso H., Sousa-Silva R., Carvalho P., Martins B. and Won M. (2022b). Annotating arguments in a corpus of opinion articles. In *Proceedings of the Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 1890–1899.

Rodrigues F., Pereira F. and Ribeiro B. (2013). Learning from multiple annotators: Distinguishing good from random labelers. *Pattern Recognition Letters* **34**(12), 1428–1436.

Rowe G. and Reed C. (2008). Argument diagramming: The araucaria project. In Okada A., Shum S.B. and Sherborne T. (eds), *Knowledge Cartography: Software Tools and Mapping Techniques*. London: Springer London, pp. 164–181.

Ruder S. (2017). An overview of multi-task learning in deep neural networks. ArXiv, abs/1706.05098.

Ruder S., Bingel J., Augenstein I. and Søgaard A. (2019a). Latent multi-task architecture learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, Honolulu, pp. 4822–4829.

Ruder S., Vulić I. and Søgaard A. (2019b). A survey of cross-lingual word embedding models. *Artificial Intelligence Research* **65**, 569–631.

Schaefer R. and Stede M. (2021). Argument mining on twitter: A survey. *Information Technology* **63**(1), 45–58.

Schiller B., Daxenberger J. and Gurevych I. (2021a). Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics, pp. 380–396.

Schiller B., Daxenberger J. and Gurevych I. (2021b). Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*.

Schnitker S.A. and Emmons R.A. (2013). Hegel's thesis-antithesis-synthesis model. In Runehov A.L.C. and Oviedo L. (eds), *Encyclopedia of Sciences and Religions*. Dordrecht: Springer Netherlands, pp. 978–978.

Schröder F. and Biemann C. (2020). Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 2971–2985.

Schuster S., Gupta S., Shah R. and Lewis M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 3795–3805.

Sharma S., Zheng G. and Awadallah A.H. (2021). Metaxt: Meta cross-task transfer between disparate label spaces. CoRR, abs/2109.04240.

Simosi M. (2003). Using toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. *Argumentation* **17**(2), 185–202.

Skeppstedt M., Peldszus A. and Stede M. (2018). More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, Brussels, Belgium. Association for Computational Linguistics, pp. 155–163.

Snow R., O'Connor B., Jurafsky D. and Ng A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii. Association for Computational Linguistics, pp. 254–263.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Spliethöver M. and Wachsmuth H. (2020). Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, Online. Association for Computational Linguistics, pp. 76–87.

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

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

Stab C., Miller T., Schiller B., Rai P. and Gurevych I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 3664–3674.

Stede M. and Schneider J. (2018). *Argumentation Mining*. Morgan & Claypool Publishers.

Swales J.M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Teufel S. (1998). Meta-discourse markers and problem-structuring in scientific articles. In Discourse Relations and Discourse Markers.

Thomas S.N. (1986). *Practical Reasoning in Natural Language*, 3rd Edn. Englewood Cliffs, NJ: Prentice Hall.

Tjong Kim Sang E.F. and Veenstra J. (1999). Representing text chunks. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway. Association for Computational Linguistics, pp. 173–179.

Toulmin S.E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

van Eemeren F.H. (2001). *Crucial Concepts in Argumentation Theory*. Amsterdam: Amsterdam University Press.

van Eemeren F.H. (2018). *Argumentation Theory: A Pragma-Dialectical Perspective*. Cham: Springer Verlag.

van Eemeren F.H., Garssen B., Krabbe E.C.W., Snoeck Henkemans A.F., Verheij B. and Wagemans J.H. (2014). *Handbook of Argumentation Theory. Springer Reference*. Dordrecht: Springer Netherlands.

van Eemeren F.H. and Grootendorst R. (1984). *Speech Acts in Argumentative Discussions*. Berlin, New York: De Gruyter Mouton.

van Eemeren F.H. and Grootendorst R. (1987). Fallacies in pragma-dialectical perspective. *Argumentation* **1**(3), 283–301.

van Eemeren F.H. and Grootendorst R. (2004). *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press.

van Eemeren F.H., Houtlosser P. and Snoeck Henkemans A.F. (2007). *Argumentative Indicators in Discourse: A Pragma-Dialectical Study*. Dordrecht: Springer.

van Eemeren F.H., Jackson S. and Jacobs S. (2015). Argumentation. In *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*. Cham: Springer International Publishing, pp. 3–25.

van Gelder T. (2001). The reason! project. *The Skeptic* **21**(2), 9–12.

Visser J., Budzynska K. and Reed C. (2017). A critical discussion game for prohibiting fallacies. *Logic and Logical Philosophy* **27**(4), 491–515.

Visser J., Konat B., Duthie R., Koszowy M., Budzynska K. and Reed C. (2020). Argumentation in the 2016 US presidential elections: Annotated corpora of television debates and social media reaction. *Language Resources and Evaluation* **54**, 123–154.

Visser J., Lawrence J., Reed C., Wagemans J. and Walton D. (2021). Annotating argument schemes. *Argumentation* **35**(1), 101–139.

Visser J., Lawrence J., Wagemans J. and Reed C. (2018). Revisiting computational models of argument schemes: Classification, annotation, comparison. In Modgil S., Budzynska, K., Lawrence, J. and Budzynska, K. (eds), *Computational Models of Argument - Proceedings of COMMA 2018*. Frontiers in Artificial Intelligence and Applications, vol. 305, Netherlands. IOS Press, pp. 313–324.

Wachsmuth H., Naderi N., Hou Y., Bilu Y., Prabhakaran V., Thijm T.A., Hirst G. and Stein B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain. Association for Computational Linguistics, pp. 176–187.

Wagemans J. (2016a). Constructing a periodic table of arguments. In *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pp. 1–12.

Wagemans J. (2019). Four basic argument forms. *Research in Language* **17**, 57–69.

Wagemans J.H.M. (2016b). Argumentative patterns for justifying scientific explanations. *Argumentation* **30**(1), 97–108.

Walker M., Tree J.F., Anand P., Abbott R. and King J. (2012). A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA), pp. 812–817.

Walton D. and Macagno F. (2015). A classification system for argumentation schemes. *Argument & Computation* **6**(3), 219–245.

Walton D., Reed C. and Macagno F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.

Walton D.N. (1996). *Argumentation Schemes for Presumptive Reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Walton D.N. (2006). *Fundamentals of Critical Argumentation. Critical Reasoning and Argumentation*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1351324923000062 Published online by Cambridge University Press

Wang X., Shi W., Kim R., Oh Y., Yang S., Zhang J. and Yu Z. (2019). Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 5635–5649.

Wegerif R. (2008). Dialogic or dialectic? the significance of ontological assumptions in research on educational dialogue. *British Educational Research Journal* **34**(3), 347–361.

Wyner A., Mochales-Palau R., Moens M.-F. and Milward D. (2010). Approaches to text mining arguments from legal cases. In Francesconi E., Montemagni S., Peters W. and Tiscornia D. (eds), *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 60–79.

Wyner A., Schneider J.A., Atkinson K. and Bench-Capon T. (2012). Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument - Proceedings of COMMA 2012*. Frontiers in Artificial Intelligence and Applications, vol. 1, USA. IOS Press, pp. 43–50.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R.R. and Le Q.V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E. and Garnett R. (eds), *Advances in Neural Information Processing Systems*, vol. 32, Vancouver, Canada. Curran Associates, Inc.

Ye L.R. and Johnson P.E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly* **19**(2), 157–172.

Zhang G., Nulty P. and Lillis D. (2022). A decade of legal argumentation mining: Datasets and approaches. In Rosso P., Basile V., Martínez, R., Métais E. and Meziane F. (eds), *Natural Language Processing and Information Systems*. Cham: Springer International Publishing, pp. 240–252.

Zukerman I., McConachy R. and George S. (2000). Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, Mitzpe Ramon, Israel. Association for Computational Linguistics, pp. 55–62.

Cite this article: Lopes Cardoso H, Sousa-Silva R, Carvalho P and Martins B (2023). Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering* **29**, 1150–1187. https://doi.org/10.1017/S1351324923000062