# Argumentation and Domain Discourse in Scholarly Articles on the Theory of International Relations

**Magdalena Wolska*** **Sassan Gholiagha†** **Mitja Sienknecht†**

**Dora Kiesel*** **Irene Lopez Garcia*** **Patrick Riemann‡** **Matti Wiegmann***

**Bernd Fröhlich*** **Katrin Girgensohn†** **Jürgen Neyer†** **Benno Stein***

*Bauhaus-Universität Weimar †Europa-Universität Viadrina ‡Jönköping University
firstname.lastname@{uni-weimar.de|europa-uni.de|ju.se}

## Abstract

We present the first dataset, an annotation scheme, discourse analysis, and baseline experiments on argumentation and domain content types in scholarly articles on political science, specifically on the theory of International Relations (IR). The dataset comprises over 1 600 sentences stemming from three foundational articles on Neo-Realism, Liberalism, and Constructivism. We show that our annotation scheme enables educationally-relevant insight into the scholarly IR discourse and that state-of-the-art classifiers, while effective in distinguishing basic argumentative elements (Claims and Support/Attack relations) reaching up to 0.97 micro $F_1$, require domain-specific training and fine-tuning on the more fine-grained tasks of relation and content type prediction.

## 1 Introduction

While most prior research into the universe of political discourses has been focused on the genres of debate and speeches, studies of academic political discourse have been lacking. One of the goals of the SKILL project, from which this paper stems, is to fill this gap.* Our goal is to develop and apply AI technologies to facilitate analysis of argumentation in scholarly articles in political science, especially in the context of teaching the theory of International Relations (IR). Since political discourse has been shown to be complex and intertextual (Chilton and Schäffner, 2002), we would like to provide students with AI tools which would facilitate comprehension of original articles used in teaching syllabi and coach them in identifying and producing expert argumentation in the field.

In order to gain insight into the structure and properties of arguments in the domain of IR theory, we developed an annotation scheme which enables analysis of scholarly IR discourse in terms of interaction between argumentation and types of domain content contributing to arguments. The scheme comprises two orthogonal dimensions: discourse and content domain. At the discourse dimension, we model argumentation (basic premise-conclusion structures) and basic rhetorical structure (elaborative discourse segments). At the domain dimension we focus on contributions relevant from the point of view of the IR domain and distinguish between theoretically and empirically-focused statements. We apply the scheme to three major theory-foundational articles in IR and address the following research questions relevant from the point of view of teaching IR based on these sources using NLP: (RQ1) To what extent is evidence for claims explicitly provided in the texts? (RQ2) Is argumentation mainly grounded in theory or in empirics? (RQ3) Can state-of-the-art language models reliably identify the basic argumentation elements?

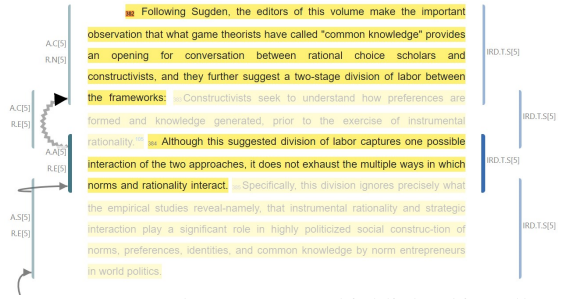


Figure 1: Annotated sentences are highlighted in yellow. Discourse (on the left) and domain (right) categories are shown with their confidence. Line arrows are support relations; zig-zagged attack relation is in focus.

Our contributions are: (1) an analytical model for annotating theory-oriented political discourse, (2) an expert-annotated corpus of three foundational IR papers, (3) an analysis of premise-conclusion structures in this corpus, and (4) baseline computational models for identifying basic argumentation elements in scholarly IR discourse.

*SKILL stands for "Sozialwissenschaftliches KI-Labor für Forschendes Lernen" (en. A social science lab for research-based learning)



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 9238–9249
January 19–24, 2025. ©2025 Association for Computational Linguistics

## 2 Related Work

Numerous studies have focused on the analysis of various aspects of political discourse, including modeling political debates (Vilares and He, 2017; Haddadan et al., 2019; Padó et al., 2019; Goffredo et al., 2022; Mancini et al., 2022), creation of corpora such as the DCEP (Hajlaoui et al., 2014) or JRC-Acquis (Steinberger et al., 2006) and tagged corpora of parliamentary debates (see, for instance, (Abercrombie and Batista-Navarro, 2018, 2020), studies of parliamentary language based on national parliament corpora (Chilton, 2004; Bayley, 2004), analysis of specific political speeches (Beelen et al., 2017; Labbé and Savoy, 2021; Card et al., 2022), or analysis of higher-level pragmatic aspects such as bias (Fischer-Hwang et al., 2020; Davis et al., 2022), manipulation, and politeness (Abuelwafa, 2021; Moghadam and Jafarpour, 2022; Kádár and Zhang, 2019; Trifiro et al., 2021).

Related work specific to annotating argumentation is included directly where we discuss design choices for our annotations scheme (see Section 5).

## 3 Data

### 3.1 Corpus Selection

As a basis for modeling the discourse of International Relations we selected three scientific articles, each representative of one of the mainstream theories in the field of IR, namely Neo-Realism (Waltz, 1993) (further referred to as Waltz), Liberalism (Putnam, 1988) (Putnam), and Constructivism (Finnemore and Sikkink, 1998) (Finnemore). The selection and focus on the three theories are due to their relevance for the discipline as such and for the differentiation of the theories themselves. The selected texts are considered central in the respective theoretical tradition. The articles were selected on the basis that they i) introduce the basic building blocks of the theory, ii) make fundamental theoretical statements, and iii) are considered representative of their field. In addition, these are the central texts that are on the syllabus in many introductory courses on IR theories. In this way, we guarantee the empirical relevance of the text selection for political science students. Each of the articles covers 20-30 pages of text. Table 1 shows basic descriptive information about the corpus.

### 3.2 Preprocessing

The corpus was prepared by segmenting into sentences in a semi-automatic process. A sentence was defined, in a standard fashion, as a linguistic unit which expresses a complete thought and typically consists of a subject and predicate. Aside from the typical end-of-sentence punctuation (full-stop, question mark, and exclamation mark), sentence boundaries were also identified by semicolons, colons and (em) dashes which are often used in scholarly articles to delimit clauses that could also be rendered as separate sentences; semicolon-delimited fragments were not split from their superordinate clauses. Sentence segmentation was verified and corrected manually by one of the co-authors with linguistic background. Only the body of the articles—without footnotes and endnotes—was used for annotation and further analysis.

| Statistic | Waltz | Putnam | Finnemore | Total |
|---|---|---|---|---|
| Sentences | 640 | 466 | 513 | 1 619 |
| Tokens | 13 069 | 11 799 | 13 042 | 37 910 |
| Types | 2 762 | 2 544 | 2 536 | 5 516 |

Table 1: Corpus key figures. Number of sentences denotes also the number of annotated markables.

## 4 Methodology

### 4.1 Scheme Development

The annotation scheme (see Section 5 and Figure 2) was developed in a combined theory- and data-driven fashion by senior researchers in IR and Linguistics, co-authors of this paper.

The Discourse dimension was derived from existing approaches to rhetorical structure and argumentation analysis. Key modification was simplification: we opted for a basic model of premise-conclusion structures with the view to obtaining high agreement. While the initial set of rhetorical functions included also, for instance, (rhetorical) Questions and Quotations, we ultimately did not ask annotators to code them since these categories can be reliably identified semi-automatically.

The Domain dimension was developed over several iterations of alternating bottom-up and top-down attempts for which we used Waltz's earlier article (Waltz, 1988) as development data. Tentative Domain subschemes included more fine-grained categories and alternative definition wording.

We started with three categories at the level of theoretical statements intended to model the world within IR theories directly: Assumptions (basic underlying ideas of a theory about how the world works), Processes (statements about dynamics or causal mechanisms in the world which are sub-

ject to Assumptions), and Outcomes (products of Processes). These proved not fine-grained enough: most markables obtained multiple categories. We therefore differentiated between three types of theoretical statements: Foundational, Assumption, and Inference. Foundational statements were meant as building blocks shared between IR theories, Assumptions as statements laying out specific theory's premises, and Inferences as derivable from either Foundational statements or Assumptions. Inferences were marked with attributes as to whether they are about entities (Actors/Structures) or related to events (Processes/Outcomes); not shown in Figure 2. The distinction between Assumptions and Inferences proved difficult to pinpoint rigorously, which led to a large disagreements between annotators, whereas attributes proved difficult due to their dual nature: some events can be viewed as processes or outcomes depending on perspective.

Empirically-focused statements were initially subcategorized into Counterfactual (alternative past scenario), Hypothetical (possible future scenario), and Factual (actual historic/present event(s)). The distinction between the former two proved difficult, thus we introduced a broader category for Speculative statements and Evaluative as the category for presenting or evaluating world events from the perspective of a theory or the author's position.

Ultimately, we arrived at a scheme that is a compromise between reliability, cost, and descriptive power: our annotators reach satisfactory agreement on a model that targets theory and argumentation-oriented research questions of our interest.

### 4.2 Annotation Procedure

**Annotators** Annotation was carried out by two senior domain experts and student researchers. The former guided the process and were responsible for setting the gold standard. Central to the selection of annotators was, first, their domain knowledge, reflected in understanding of IR theories. This knowledge was assessed in interviews and documented based on attended courses. Second, a good command of English. Third, a willingness to familiarize themselves with machine-aided text analysis. Fourth, a high level of reliability, independence, commitment and ability to work in a team.

All hired annotators are studying in a Political Science program. At the start of employment, 7 out of 11 annotators had a Bachelor's degree in a relevant field (Political Science, Politics and Law, Cultural Studies). All of them were primarily educated in Western European universities. Four annotators are female, and six are male. Some fluctuation occurred in the group of student annotators, but the majority of them have been with the project from the start. Those who joined later have been onboarded and trained rigorously. At the time of writing, no discernable effect on annotation quality can be derived from the time of employment.

**Annotation Tool** We developed our own, particularly tailored, annotation software. Existing tools such as Brat, Label Studio, etc. did not meet our requirements especially in terms of text length to annotate (papers often with more than 30 pages), paper and markable management, flexibility towards hierarchical coding books or annotation schemes, and central data storage. The system is realized as a self-hosted web-application with a database containing annotations and logs, a backend for data management and analytics routines that serve different kinds of web pages suited for different aspects of the annotation process: managing users and their different roles (annotators, analysts, experts, etc.), handling of multiple papers and their markable definitions, as well as assigning annotators to their respective tasks, which could be entire papers or only selected parts of it.

For the annotation interface itself, our motivation was providing a visual appearance that mimics the look and feel of a typical scientific paper (Figure 1) and consists of all its structural elements. It also offers a special mode for self-control once the task is entirely finished. For the supervisors (domain experts) and administrators, certain information and analysis pages help to track the progress and analyze the results of the annotation tasks.

**Annotation Process** The annotation of the corpus was performed with the annotation tool by two domain experts and non-experts trained in the course of the project. Gold standard annotations were obtained from the expert annotations via disagreement mitigation: Cases of disagreement were discussed by domain experts and the project linguist until a consensus was reached.

Non-expert annotators received written annotation guidelines, were systematically trained in IR theories and category definitions, and were supervised by domain experts and the linguist. Prior to annotating an article annotators were offered workshops on background knowledge needed to understand the theory represented in given article. Annotation quality was monitored as follows: Once

coding was completed by all annotators, gold standard was released and annotators were able to compare their performance against it using the annotation tool's visual comparison mode. Annotators were asked to meet in groups of two to three to discuss disagreements and to prepare a list of questions on markables whose gold standard annotation was unclear; these were then sent to the supervisors. Supervisors met once a week to discuss group questions and prepare explanations. The annotators and supervisors then met (also once a week) to clarify questions and discuss specific individual markables. The annotators received feedback on their individual performance in one-on-one meetings, where patterns of annotator-specific deviations from the gold standard were also pointed out.

## 5 Results

### 5.1 A Model of IR Theory Discourse

Our model of scholarly IR discourse comprises two dimensions, discourse and domain. The **discourse dimension** describes argumentation and the rhetorical structure and can be applied to any argumentative text. The **domain dimension** describes the discourse contributions in terms of the type of content specific to the domain of discourse; in our case, the theory of International Relations. The overview of the complete annotation scheme is shown in Figure 2. The grayed out elements represent categories annotated in a semi-automatic way and manually verified or annotated by a linguist. We do not report inter-annotator agreement for these categories, but include them here for the sake of completeness. Annotation categories within the two dimensions are defined below. Examples for each of the categories are shown in the Appendix (see Table 7).

**Discourse Dimension** The discourse dimension models argumentation and rhetorical aspects of text. At the level of **Argumentation** we model discourse structures which build up an argument, that is, we identify those discourse moves that contribute to bringing argumentation forward as well as relations between those moves. Our argumentation-related categories are a simplified subset of argumentative moves proposed by Toulmin (2003). The original Toulmin model of argumentation has been widely used in studies of argumentative discourse, however, it has been shown to present difficulties for annotation of real life argumentation (see, for instance, (Simosi, 2003)). Torsi and Morante (2018) show that argumentation is in general difficult to annotate and yields low inter-annotator agreement. We therefore opt to model argumentation at the lowest level of complexity, namely, by only identifying basic *premise-conclusion structures* in terms of Claims and two relations that may hold between them, Support and Attack, defined as follows:

**Claim** is a statement that presents a basic building block of an argument. It is the assertion that a party puts forth and would like to convince the audience of, that is, prove. A claim can be also thought of as the conclusion that a party in discourse is attempting to draw.

**Support** in an argument is a statement that provides evidence justifying a claim. This may be a statement that directly brings up facts, data, or other pieces of evidence showing why a claim holds. The purpose of a supporting statement is to increase credibility of a claim, i.e. the readers' belief that the claim holds.

**Attack** is a counter-argument to a previously proposed claim. The purpose of an attacking statement is to decrease credibility of a claim, i.e. the readers' belief that the claim holds.

Note that unlike other argumentation annotation schemes (e.g. (Stab and Gurevych, 2014; Peldszus and Stede, 2015)) we do not distinguish between so-called main/major and minor claims at this point. We refrain from adding to the complexity of annotation since our data comprises research articles, i.e., longer discourses of high linguistic complexity. However, we approximate the distinction between major and minor claims by modeling local elaboration structures at the rhetorical level explicitly (see below). Discourse units which are not argumentative in the sense of the three categories defined above remain unlabelled at the argumentative level.

At the level of **Rhetorical Moves** we model the structural organization of text, i.e. the rhetorical roles of spans of text in a larger discourse which make the discourse coherent. Depending on a linguistic theory, rhetorical phenomena in discourse may encompass up to even 30 types of rhetorical coherence types (Taboada and Mann, 2006) including relations such as Background (facilitates understanding), Evaluation (evaluative comment), Purpose (intent behind a state or action), Means (method or instrument that facilitates realization of an action). Note that argumentation itself is also a rhetorical phenomenon which can be modelled at finer detail than the Claim-Support/Attack

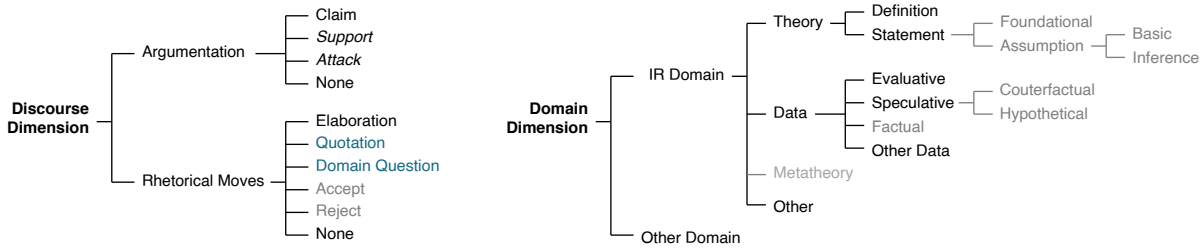


Figure 2: Overview of the annotation scheme. Support and Attack are *relations* between Claims. Quotation and Domain question (in blue) are semi-automatically derived. Labels in gray complete the schema, but are not acquired from the annotators (see Section 4.1). "None" is a technical category where no construct applies at this level.

model presented above, using rhetorical relations such as Evidence, Explanation, and (volitional and non-volitional) Cause and Result. We model argumentation as a distinct level of annotation since it pays a central role in our model and we focus directly on basic argumentative premise-conclusion structures. At the rhetorical level we annotate a single relation, Elaboration, defined as follows:

**Elaboration** expands on a point by contextualizing it or provides more information about a previous statement. It may describe it in a different way (e.g. restate, paraphrase, or reformulate it) or at a different level of abstraction (e.g. make it more specific/general)

Elaboration as defined above combines presentational aspects (cf. Mann and Thompson's (1988) Reformulation/Restatement and Summary and Hobbs' (1979) Repetition) as well as content aspects (cf. Danlos' (2001) Particularization and Generalisation of a discourse unit. The main purpose of Elaboration in our scheme is to facilitate setting apart main claims from minor (elaborated) claims in argumentation. Non-elaborative statements remain unannotated at the rhetorical level.

**Domain Dimension** The types of content contributed to discourse depends on the discourse genre and, naturally, on the domain of discourse. For instance, in the medical domain there might be discourse contributions related to a patient's diagnosis, in the music domain to the structure of a musical piece, and in the domain of chemistry to the interactions between chemical elements. In our case of political science domain, the domain dimension models the type of content specific to presenting a political science theory, in particular, theory of International Relations. For statemens within the **IR Domain**, that is those about International Relations or global politics, we distinguish between content related to Theory and Data with two subtypes defined as follows:

**Theory** statements present theoretical postulates. There may be empirical references or illustrations within theory statements, however, as soon as a theoretical assertion is presented as a generalization beyond any specific empirical references or illustrations, it should be annotated within the Theory category. Two subcategories are explicitly defined:

**Definition** is a statement which explicitly specifies a meaning of a term used in the domain.

**Theoretical Statement** is a non-definitional theoretical statement, i.e. one which is about IR-relevant theoretical concepts or topics.

**Data** statements provide relevant empirical evidence, that is, concrete reference to the real world, including classes of events (e.g. war), or social facts. Two subcategories are explicitly defined:

**Speculative** makes a statement about a possible present or future scenario or an alternative past scenario; neither has actually happened.

**Evaluative** contains a reference to real world events, data, or (social) facts which are evaluated or interpreted by the author from any theoretical standpoint or presents it as a fact through a theory's perspective.

Any other statements about the real world annotated as **Data Other**.

If a statement does present IR relevant content which cannot be classified as Theory or Data according to definitions above, we annotate it as **IR Theory Other**; this makes the scheme open-ended at the domain dimension. If a statement explicitly refers to a domain other than political science, IR, or global politics, it is annotated as **Other Domain**. Figure 2 shows the category structure of the annotation scheme and Figure 1 illustrates an annotated excerpt from Waltz's 1993 article.

| Categories | Inter-annotator agreement (n ; $\kappa$ (SE) ; PABAK) | | |
|---|---|---|---|
| | **Waltz** | **Putnam** | **Finnemore** |
| Claim \| None | 639 ; 0.781 (0.051) ; 0.947 | 468 ; 0.717 (0.090) ; 0.962 | 512 ; 0.495 (0.089); 0.906 |
| IR Domain \| Other Domain | 588 ; — ; 0.997 | 445 ; — ; 1 | 475 ; 0.635 (0.087) ; 0.937 |
| Theory \| Data \| Other | 587 ; 0.748 (0.028) ; 0.792 | 445 ; 0.659 (0.035) ; 0.685 | 446 ; 0.546 (0.040) ; 0.605 |
| State \| Def \| Eval \| Spec \| Misc | 526 ; 0.865 (0.020) ; 0.844 | 375 ; 0.909 (0.021) ; 0.915 | 345 ; 0.936 (0.023) ; 0.936 |

Table 2: Inter-annotator agreement between domain experts on Discourse and Domain categories measured by Cohen's kappa and PABAK. $n$ is the number of observations in the given subset of ubcategories. For subsets exhibiting an extreme prevalence problem $\kappa$ value is not shown.

### 5.2 Annotation Quality

We performed an inter-annotator reliability analysis to assess the degree that annotators consistently assigned Discourse and Domain categories to the markables in the corpus.[†] Marginal distributions indicated the prevalence problem in the IR Domain vs. Other Domain subcategory. The problem did not occur in any of the other categories suggesting that Cohen's $\kappa$ (Cohen, 1960) is an appropriate measure and PABAK was computed for the problem categories (Eugenio and Glass, 2004).

Cohen's kappa was computed to assess the agreement between the two domain experts and between each of the non-expert annotators and the gold standard in assigning Discourse and Domain categories to the markables in the corpus. We computed results for 6 non-expert annotators who completed annotation of all the articles in the corpus (fully-crossed). Note that our annotation scheme has a hierarchical structure. This means that error on higher categories results in propagated error on lower categories. As in previous studies which used coding schemes with dependent categories—see, for instance, (Ruiter et al., 2022)—we exclude the propagated error and for each category calculate $\kappa$ only on the subset of instances on which there is an agreement on the higher category.

Detailed inter-annotator agreement results for experts are shown in Table 2. Agreement between expert annotators ranged from $\kappa = 0.45$ (moderate) to $\kappa = .94$ (almost perfect agreement) (Landis and Koch, 1977), with perfect agreement on categories with prevalence problem. Low agreement occurred only on argumentation categories and is not surprising since argumentation has been shown to be difficult to annotate and the result is in line with previously published agreement estimates on coding similar constructs. For non-expert annotators averaged agreement ranged from $\kappa = 0.36$ (fair) to $\kappa = .79$ (substantial). The additional difficulty with our data for non-experts lies in the fact that strong and broad background domain knowledge is required—not only in IR theory, but also in general political history—in order to comprehend and annotate our domain. The substantially higher agreement between experts than between non-experts and the gold standard reflects this.

### 5.3 Argumentation in IR Discourse

The annotated gold standard corpus comprises three articles—1619 sentences in total—coded with the categories defined in Section 5.1. Table 3 shows basic descriptives on the categories split by article.

The total number of Claims in the corpus is 1546. There are 701 Claim-Support and 43 Claim-Attack pairs. As expected, all texts are for the most part argumentative with less than 10% sentences without argumentative function. The majority of Claims in all three articles are part of elaborated structures, i.e. between 22 and 31% of all segments are what can be considered "main" or "major" claims, i.e. they possibly initiate an elaborated segment and can be thought of as the core train of reasoning.

**Supported vs. Unsupported Claims (RQ1)** In order to answer the first research question as to whether claims are justified (see RQ1 in Section 1), we look into the proportion of supported and unsupported claims.

Only 178 out of the 598 Claims in Waltz's text (around 30%), 156 in Putnam's (around 34%), and 210 in Finnemore's & Sikkink's (around 42%) are provided with supporting evidence within text, that is, they form Claim-Support chains and can be considered arguments in the sense of Premise-Conclusion structures. The majority of those are provided with a single evidence statement (131, 115, and 165, respectively). Waltz provides up to 6 Supports for 12 Claims, 41 Claims in Putnam's text have multiple Supports (one Claim with 8 Supports), and 45 in Finnemore's & Sikkink's text (up

[†]Markables with erroneously missing annotations were excluded from the analysis.

| Statistic | Waltz | Putnam | Finnemore |
|---|---|---|---|
| *Argumentation* | | | |
| Claim | 598 (0.93) | 454 (0.98) | 494 (0.96) |
| Support | 231 (0.36) | 204 (0.44) | 266 (0.52) |
| Attack | 18 (0.03) | 12 (0.03) | 13 (0.03) |
| None | 42 (0.07) | 14 (0.03) | 19 (0.04) |
| *Rhetorical Moves* | | | |
| Elaboration | 439 (0.69) | 344 (0.74) | 399 (0.78) |
| Other | 201 (0.31) | 124 (0.26) | 114 (0.22) |
| *IR Domain* | | | |
| Theory Statement | 198 (0.31) | 294 (0.63) | 324 (0.63) |
| Theory Definition | – | 11 (0.02) | 7 (0.01) |
| Data Evaluative | 305 (0.48) | 122 (0.26) | 77 (0.15) |
| Data Speculative | 89 (0.14) | 19 (0.04) | 1 (0.0) |
| Data Other | 5 (0.01) | 1 (0.0) | 4 (0.01) |
| Other | 43 (0.07) | 7 (0.01) | 59 (0.12) |

Table 3: Markables split by type in each of the three articles. Shown are the number of markables by type and fraction of all markables within that article.

to 4 Supports). The remaining Claims are presented without evidence in the text.

It can be assumed that the authors consider justification for the unsupported Claims to be part of the so-called common ground or shared understanding/common knowledge, i.e. "general knowledge shared by the speaker, hearer, and audience" (Walton, 1996); see also (Clark and Schaefer, 1989; Van Eemeren et al., 2004). While in itself the fact that many Claims are provided without Support is not surprising, it has implications on teaching IR theories based on these articles: the possible background knowledge gaps need to be filled in.

**Theory vs. Data-driven Argumentation (RQ2)** In general, Putnam's and Finnemore's & Sikkink's texts are more theoretically oriented (around 67% of all Claims in the Theory category) whereas Waltz's text more empirically (around 66% of all Claims in the Data category). Interestingly, as far as type of supporting evidence provided, the majority of supporting statements in Waltz are based in empirical Data (72%), in Finnemore in Theory (65%), wheras Putnam's argumentation refers equally to Data and Theory (51% of Supports are Theory).

From an educational perspective this means that the key prerequisites for comprehending Waltz's argumentation are strong background in history and an ability to recognize the impact of world events on international relations, whereas for comprehending Finnemore's argumentation strong background in the theory of IR is required. Putnam's text requires knowledge of both. Instructors who use original sources as part of undergraduate syllabi should be aware of those requirements.

| Split | is claim | | has relation | | is support | |
|---|---|---|---|---|---|---|
| | pos | neg | pos | neg | pos | neg |
| Training | 641 | 655 | 544 | 544 | 552 | 33 |
| Validation | 80 | 82 | 68 | 68 | 69 | 4 |
| Test | 81 | 82 | 68 | 68 | 70 | 5 |

Table 4: Sizes of the evaluation dataset. We sampled three datasets from the documents in the corpus and split each 80:10:10 into training, validation, and test.

### 5.4 Computational Modeling

To scale the analysis of argumentation in political discourse and study argumentation synthesis, we need to develop computational models for our annotation scheme. Hence, we conduct a quantitative evaluation of the annotations through a series of three typical argument mining experiments. All tasks are binary classification and the annotations of the argumentative domain can be fully constructed by solving all three tasks consecutively.

First, **claim detection** aims to decide if a markable is a claim or not. We collected all markables annotated as a claim across all three documents as the positive class, and all others (support, attack, and non-argumentative) as the negative class. The resulting dataset is nearly balanced (cf. Table 4).

Second, **relation prediction** aims to decide if there is a relation between a claim and any other non-claim markable in the same paragraph. We constructed a dataset of (claim, candidate premise)-tuples from all three documents by selecting a claim and pairing it with all non-claim markables from the same paragraph. A tuple was assigned to the positive class if the candidate markable supports or attacks the claim and to the negative class otherwise. We balanced this dataset by randomly under-sampling the negative class.

Third, **Support/Attack classification** aims to decide if, given a (claim, premise)-tuple, the premise is a Support or Attack of the claim. We constructed a dataset by extracting all annotated (claim, premise)-tuples from the three documents. A tuple was assigned to the positive class if the existing relation was Support and to negative class otherwise. This dataset is very imbalanced since Attack relations are scarce. It should be noted that relation prediction and Support/Attack classification are often combined in the related work, however, we split those tasks to be comparable to the IBM Project Debater API[‡].

[‡]https://developer.ibm.com/apis/catalog/debater--project-debater-service-api/

| Model | Task | Accu. | Micro $F_1$ | Binary $F_1$ | |
|---|---|---|---|---|---|
| | | | | Positive | Negative |
| BERT | is claim | 0.71 | 0.71 | 0.73 | 0.70 |
| BERT | has relation | 0.62 | 0.62 | 0.62 | 0.61 |
| BERT | is support | 0.97 | 0.97 | 0.99 | 0.80 |
| Debater | is support | 0.80 | 0.80 | 0.88 | 0.29 |

Table 5: Results of the RoBERTa and IBM Debater baseline experiments on three argument mining tasks.

| Statistic | is claim | has relation | | is support | |
|---|---|---|---|---|---|
| | | Claim | Premise | Claim | Premise |
| *Rhetoric* | | | | | |
| Elaboration | 112 (0.32) | 102 (0.37) | 117 (0.41) | 47 (0.21) | 71 (0.20) |
| *IR domain* | | | | | |
| Theory statement | 78 (0.30) | 81 (0.44) | 66 (0.36) | 50 (0.18) | 40 (0.20) |
| Data evaluative | 49 (0.20) | 35 (0.31) | 48 (0.46) | 13 (0.23) | 24 (0.25) |
| Data speculative | 13 (0.31) | 8 (0.38) | 7 (0.29) | 7 (0.43) | 7 (0.14) |

Table 6: Example count and Misclassification rate in the test dataset, split by *Rhetorics* and *IR Domain* categories (only BERT baselines). Categories with fewer than 10 examples are excluded.

**Models and Training** For all three tasks, we trained a transformer-encoder classification model BERT, using HuggingFace's implementation of `roberta-base` with a pooling-based classification head, AdamW optimizer, and linear learning rate decay. For each model, we conducted a parameter grid-search over the learning rates `(1e-5, 2e-5, 5e-5)` and the epoch sizes `(10, 15, 20)` on the validation split, and tested on the best performing configuration respectively for each task.

In addition, we also used the pro-con endpoint of the IBM Project Debater API for Support/Attack classification. Since the API does not need to be trained, we evaluated the complete dataset (not just the test split). This means the results are not strictly comparable to BERT, but much more reliable regarding the Attack relations. Since Debater judges a relation as pro, con, or neutral with relative scores, we applied a simple heuristic: if the pro score is larger than the con score, we rated the premise as a Support (the positive class) and vice versa.

**Classification Quality (RQ3)** We quantitatively evaluate the models via accuracy and micro-averaged $F_1$ (cf. Table 5), since the tasks are all binary classification. Additionally, we inspect the binary $F_1$, particularly for the support-attack classification, which is very imbalanced and has only a few negative examples.

First, the results show that our baseline is effective for Support/Attack classification with a micro $F_1$ of 0.97. The score is likely positively distorted by the training data imbalance, however, both Debater and BERT score high on the positive (Support) examples, and BERT scores also high on negative examples, which is non-trivial even with few examples. This suggests that Support/Attack classification can be done effectively on this data.

Second, the results show a reasonable performance of the baseline model on Claim detection with an accuracy of 0.71. This is a good result, given that the model was not given any context and the decision was often difficult for the human annotators We expect that a more sophisticated model can reach notably higher performance, especially when given the pre and succeeding context.

Third, the results show that relation prediction, with an accuracy of 0.62, is the most difficult of the three tasks. This is consistent with the observations from the annotation experiments: Annotators made more errors when deciding onrelations, in most cases missing Support relations altogether; one annotator, on the other hand, overgenerated Supports. We expect that adding context can improve the performance slightly but not substantially.

Additionally, we evaluate the misclassification rate of all three tasks split by *Rhetorical Moves* and *IR Domain* markable type (cf. Table 6). There are a few instances where markables of a certain type had a notably outlying misclassification rate. First, in Claim detection, *Data Evaluative* is less often misclassified (by 0.1–0.12) than other types. Second, in Support/Attack classification, *Data Speculative* as a Claim is much more often misclassified (by 0.2–0.25). As of now, we have no convincing explanations for these outliers other than the fact that sentences in the Data category are in general lexically diverse in their wording both within and between theories since they refer to real world events which, aside from Putnam's case example brought up throughout the article, may be each time different; *Data Speculative* is also a spase category in the data (only 109 instances). Third, there is a large spread in misclassification rates within relation prediction, which might be due to the general poor performance of the model on the task.

## 6 Conclusions

In this paper we develop an analytical model for annotating theory-oriented articles on International Relations. We apply this model to three foundational articles on IR theories and show that satis-

factory agreement among domain experts can be reached. Our non-expert annotators were capable of reaching modest agreement with the gold standard on the argumentative dimension of the scheme (in line with prior work on annotating argumentation), however, the domain dimension proved more stable with agreements within the moderate range. To our knowledge our corpus is the first annotated dataset in the domain of scholarly political science discourse.[§]

Unlike in prior work in NLP, our annotation scheme and analysis links argumentation to domain content in our domain. We show that exploring the interaction between argumentative discourse aspects and its domain-specific aspects enables drawing application-relevant conclusions; in our case, identifying implications for teaching based on original sources at the university level. We also perform baseline machine learning experiments using state-of-the-art models based on our data, showing that while important argumentative aspects can be readily learned, the combination of domain content types and argumentation remains challenging.

## Limitations

We see three limitations of this work. First, the corpus includes only three articles from three theories of International Relations. While the papers we selected are indeed foundational for the three theories and and of great importance for teaching theories, texts from other theories—e.g., feminism—would need to be included. Nevertheless, in terms of number of markables our corpus is of comparable or larger size to other argumentation annotated corpora frequently exploited by the NLP community such as (Stab and Gurevych, 2014; Peldszus and Stede, 2015). What sets our corpus apart is also the fact that unlike those corpora our data includes annotation at the domain level enabling analysis of interactions between these two dimensions.

Second, we applied the annotation scheme to only one article *type* in International Relations research, since our focus in this work has been on theory-oriented papers. Other types of work in the domain—e.g., quantitative research, case studies—might require augmenting the model of content types in the scheme.

Third, we are not sure how the models (both analytical and computational) would perform on texts from other authors. Already informal analysis of the documents in our corpus showed differences in writing in terms of broadly understood "style" between the three scholars. In general, we are planning to follow up on the present work with a larger annotation project involving a large corpus of publicly available articles on International Relations.

[§]The corpus will be available for research purposes.

## Ethical Considerations

This project followed a general code of good practice regarding annotation to the best extent possible. While demographic diversity among annotators was not controlled for, the hiring opportunity was open to all candidates from the Political Science department at one of the universities participating in the project. Gender balance among annotators was ensured in the hiring process. Prior to hiring, candidates were informed about the nature of the task, the fact that their performance will be monitored, and asked to sign an Informed Consent Form concerning computational tracking of the annotation process. Annotators were systematically reminded of the quality-over-speed preference. While deadlines for assigned tasks were imposed, work was not timed to speed. Since annotating complex discourse in a highly specialized, non-trivial domain is a cognitively demanding task, annotators were also offered a weekly jour fixe with a researcher not directly involved in this work whose task was to obtain feedback on annotators' concerns and pass them on, anonymized, to the supervisors. The supervisors addressed the concerns to the best extent possible.

## References

Gavin Abercrombie and Riza Batista-Navarro. 2018. A sentiment-labelled corpus of hansard parliamentary debate speeches. In *Proceedings of ParlaCLARIN. Common Language Resources and Technology Infrastructure (CLARIN)*.

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2020. Parlvote: A corpus for sentiment analysis of political debates. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078.

Marwa Adel Abuelwafa. 2021. Legitimation and manipulation in political speeches: a corpus-based study. *Procedia Computer Science*, 189:11–18.

Paul Bayley. 2004. Cross-cultural perspectives on parliamentary discourse. *Cross-Cultural Perspectives on Parliamentary Discourse*, pages 1–390.

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, et al. 2017. Digitization of the canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3):849–864.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Paul Chilton. 2004. *Analysing political discourse: Theory and practice*. Routledge.

Paul A Chilton and Christina Schäffner. 2002. Introduction: Themes and principles in the analysis of political discourse. In *Politics as text and talk: Analytic approaches to political discourse*, pages 1–41. John Benjamins.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1).

Laurence Danlos. 2001. Event coreference between two sentences. *Computing Meaning: Volume 2*, pages 271–288.

Sara R Davis, Cody J Worsnop, and Emily M Hand. 2022. Gender bias recognition in political news articles. *Machine Learning with Applications*, 8:100304.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1).

Martha Finnemore and Kathryn Sikkink. 1998. International norm dynamics and political change. *International organization*, 52(4).

Irena Fischer-Hwang, Dylan Grosz, Xinlan Emily Hu, Anjini Karthik, and Vivian Yang. 2020. Disarming loaded words: Addressing gender bias in political reporting. In *Computation + Journalism Conference*.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 4143–4149.

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

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, Daniel Varga, et al. 2014. DCEP – Digital Corpus of the European Parliament. In *Proceedings of the 9th Language Resources and Evaluation Conference*, pages 3164–3171.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.

Dániel Z Kádár and Sen Zhang. 2019. (im) politeness and alignment: A case study of public political monologues. *Acta Linguistica Academica*, 66(2):229–249.

Dominique Labbé and Jacques Savoy. 2021. Stylistic analysis of the french presidential speeches: Is macron really different? *Digital Scholarship in the Humanities*, 36(1):153–163.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. 2022. Multimodal argument mining: A case study in political debates. In *Proceedings of the 9th Workshop on Argument Mining*, pages 158–170.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Meisam Moghadam and Niloofar Jafarpour. 2022. Pragmatic annotation of manipulation in political discourse: The case of trump-clinton presidential debate. *Linguistic Forum – A Journal of Linguistics*, 4(4):32–39.

Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanık, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Proceedings of the 1st European Conference on Argumentation*, volume 2.

Robert D Putnam. 1988. Diplomacy and domestic politics: the logic of two-level games. In *International organization*, volume 42. CUP.

Dana Ruiter, Liane Reiners, Ashwin Geet d'Sa, Thomas Kleinbauer, Dominique Fohr, Irina Illina, Dietrich Klakow, Christian Schemer, and Angeliki Monnier. 2022. Placing m-phasis on the plurality of hate: A feature-based corpus of hate online. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

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

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical papers*, pages 1501–1510.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th Language Resources and Evaluation Conference*.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Benedetta Torsi and Roser Morante. 2018. Annotating claims in the vaccination debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge University Press.

Briana M Trifiro, Sejin Paik, Zhixin Fang, and Li Zhang. 2021. Politics and politeness: Analysis of incivility on twitter during the 2020 democratic presidential primary. *Social Media + Society*, 7(3):20563051211036939.

Frans H Van Eemeren, Robert Grootendorst, and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582.

Douglas N Walton. 1996. *Argument structure: A pragmatic theory*. University of Toronto Press.

Kenneth N Waltz. 1988. The origins of war in neorealist theory like. *Journal of Interdisciplinary History*, 18(4).

Kenneth N Waltz. 1993. The emerging structure of international politics. *International security*, 18(2).

# A Examples of Categories

| | Category | Example |
|---|---|---|
| Argumentation | Claim | Normative and ideational concerns have always informed the study of international politics and are a consistent thread running through the life of International Organization. (Finnemore) |
| | Support | *Nuclear weapons do, however, narrow the purposes for which strategic power can be used.* No longer is it useful for taking others' territory or for defending one's own. (Waltz) |
| | Attack | *Some observers thought that the Spanish-American War marked America's coming of age as a great power.* But no state lacking the military ability to compete with other great powers has ever been ranked among them. (Waltz) |
| | None | What are the possibilities? (Waltz) |
| Domain | IR Domain.Theory.Definition | Voluntary defection refers to reneging by a rational egoist in the absence of enforceable contracts-the much-analyzed problem posed, for example, in the prisoner's dilemma and other dilemmas of collective action. (Putnam) |
| | IR Domain.Theory.Statement | Political action must be based on a coordination of morality and power. (Finnemore) |
| | IR Domain.Data.Evaluative | All in all, the Bonn summit produced a balanced agreement of unparalleled breadth and specificity. (Putnam) |
| | IR Domain.Data.Speculative | The negotiators might be heads of government representing nations, for example, or labor and management representatives, or party leaders in a multiparty coalition, or a finance minister negotiating with an IMF team, or leaders of a House-Senate conference committee, or ethnic-group leaders in a consociational democracy. (Putnam) |
| | IR Domain.Data.Other Data | Soldiers are trained to sacrifice life for certain strategic goals. (Finnemore) |
| | IR Domain.Other | The role of side-payments in international negotiations is well known. (Putnam) |

Table 7: Example sentences from the corpus in each category (context in italics).