SKILL-IR-Discourse: A Large, Annotated Corpus of Argumentation and Domain Discourse on International Relations

Magdalena Wolska,[1] Matti Wiegmann,[2] Sassan Gholiagha,[3] Mitja Sienknecht,[3] Dora Kiesel,[1] Irene López García,[1] Patrick Riehmann,[4] Bernd Fröhlich,[1] Katrin Girgensohn,[3] Jürgen Neyer,[3] Benno Stein[1]

[1]Bauhaus-Universität Weimar, [2]Kassel University, [3]European University Viadrina, [4]Jönköping University

firstname.lastname@{uni-weimar.de | uni-kassel.de | europa-uni.de | ju.se}

Abstract

We present a large annotated corpus of scholarly discourse in the domain of International Relations, a subfield of political science. The corpus comprises 190 articles (over 1500K tokens) annotated at the argumentation, basic rhetorical, and domain level. Five of the included articles (ca. 62K tokens) constitute a Gold-standard, coded by domain experts. The remaining articles were coded by annotators trained on the Gold-standard and monitored for annotation quality. We describe our corpus creation methodology, the annotation process and quality assurance, the corpus itself, and present insights into the data: Most argumentative structures in the data are simple premise-conclusion structures, fewer than half of the claims have explicit supporting evidence. Counter-arguments to claims are rare. The claim-to-support ratio varies widely between articles; possibly to some extent due to the topics covered (with clear common ground) or to the differences between authors' styles. The distribution of theoretical vs. evaluative statements varies strongly between articles; this can be attributed to such factors as different methodological approaches between the articles and the methodological focus of the publishing journal.

Keywords: argumentation, scholarly discourse, political science, annotated corpora

1. Introduction

Research into political discourse is largely concerned with two genres: political debates and political media. The latter includes meetings, interviews, newspaper articles, editorials, and (social) media artifacts (Randour et al., 2020). Conversely, studies of *scholarly* political discourse are scarce, albeit the history of scholarly political thought can be traced back to the Ancient Greece in the Western tradition and even earlier in the Asian political philosophy. Yet it is within scholarly discourse that new constructs are defined, political theories—models of the world—are proposed, and where theoretical reflection and argumentation takes place. With this work we contribute to filling the gap in research into understanding scholarly political discourse and construct a large corpus of scholarly political discourse on International Relations (IR) annotated at the discourse and the domain level.

In order to enable insight into the structure and properties of argumentation in IR theory, we build on an annotation scheme for analyzing this discourse genre in terms of the interaction between argumentation and types of domain content that contribute to arguments (Wolska et al., 2025). The scheme comprises two dimensions: discourse, which models basic premise-conclusion and rhetorical structures, and domain, which models claims in the IR domain as theoretically or empirically grounded. This scheme is highly detailed and requires expertise in the IR domain as well as in discourse structures, making it expensive to annotate. Case in point, the original study only annotated three articles, which is, by and large, too small to allow for generalizable insights.

We expand on this work by presenting our large scale annotation campaign, where we construct, annotate, and analyze a corpus—*SKILL-IR-Discourse*—of 190 scholarly articles (over 1,500K tokens) from major political science publishers. Five articles (ca. 62K tokens) were annotated in a rigorous fashion by domain experts, senior researchers in political science and computational linguistics. The remaining 185 articles were annotated by trained annotators, students in political science programs. To our knowledge this is the first corpus in this domain annotated at discourse and domain level at this granularity and scale.

In this paper, we describe the annotation scheme (Section 3) and describe our annotation methodology (Section 4). We also evaluate the corpus and present insights into the data (Section 5): What is the distribution of the different discourse and domain categories in political science articles and what can we learn from this distribution about the discourse in the field? Are the type frequencies stable across all documents in the corpus or do they vary? Can the variance be attributed to the text alone (stylistic differences between authors, for instance) or do external factors, such as the journal or the work's study method type, influence the variance? Finally, we also assess whether the variance is an artifact of the annotation process.

We directly release *SKILL-IR-Discourse* at: https://zenodo.org/records/17380680.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6472–6482
11–16 May 2026.

2. Related Work

Our work studies argumentation and domain statements at a granular, discourse-oriented level. This distinguishes our work from other work in the debate-oriented argumentation studies (Stab and Gurevych, 2014), such as the corpus of Lauscher et al. (2018), who annotated scientific articles in this tradition. Although political discourse in general has been analyzed extensively, our focus on political science articles, especially regarding argumentation, adds to this work. In such, our contributions relate to studies on 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). Randour et al. (2020) mention a single article analyzing academic political discourse, (Sidiropoulou, 2013), however, this work is a qualitative comparative analysis of an English to Greek translation of a coursebook—a short sample of it—and focuses on pragmatic and socio-cognitive aspects of translation in this genre. Wolska et al. (2024, 2025) present small-scale analyses of one and three, respectively, political science articles showing that the majority of claims are presented without explicit evidence and pointing at the educational implications of this asymmetry. We build on this work in that we apply Wolska et al.'s (2025) annotation scheme and validate its applicability in a large-scale study.

3. Annotation Scheme

The SKILL-IR-Discourse corpus has been annotated using the argumentation and domain discourse scheme proposed by Wolska et al. (2025). The scheme, shown in Figure 1, models the two dimensions: *Discourse* and *Domain*. The discourse dimension models *Argumentation* and *Rhetorical moves*, whereas the domain dimension models content types specific to the theory of "International Relations (IR)", a field within political science.

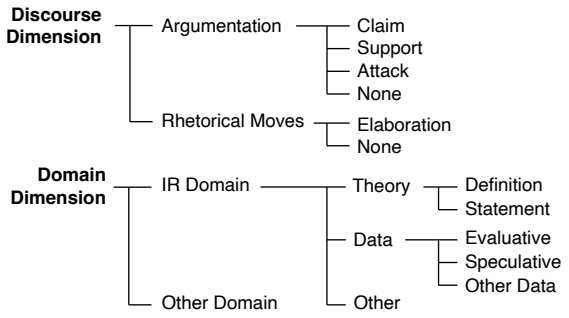


Figure 1: The annotation scheme used in this work is a reduced version of (Wolska et al., 2025) and models argumentation and domain discourse statements in two dimensions and four levels.

Argumentation The types at this level model the fundamental premise-conclusion structures of the argumentative moves of Toulmin (2003). Three categories are distinguished:

The *Claim* as the fundamental argument element. It is the assertion that is being put forth or the conclusion of an argument.

The *Support* relation between two claims, where one provides evidence to justify the other.

The *Attack* relation between two claims, where one provides a counter-argument to the other.

Rhetorical Moves Only *Elaboration* is modelled at this level. An *Elaboration* expands on a previous claim by contextualizing it or providing more information, for instance, by describing it in a different way (e.g. restating, paraphrasing, or reformulating it) or at a different level of abstraction (e.g. making it more specific or general). *Elaboration* is included in the scheme to set major claims apart from purely elaborative and thus minor claims.

Domain The types at this level model the IR-specific discourse. The scheme differentiates between data and theory.

Theory statements present theoretical postulates that are presented as generalizations beyond any specific empirical evidence (including empirical evidence within the assertion).

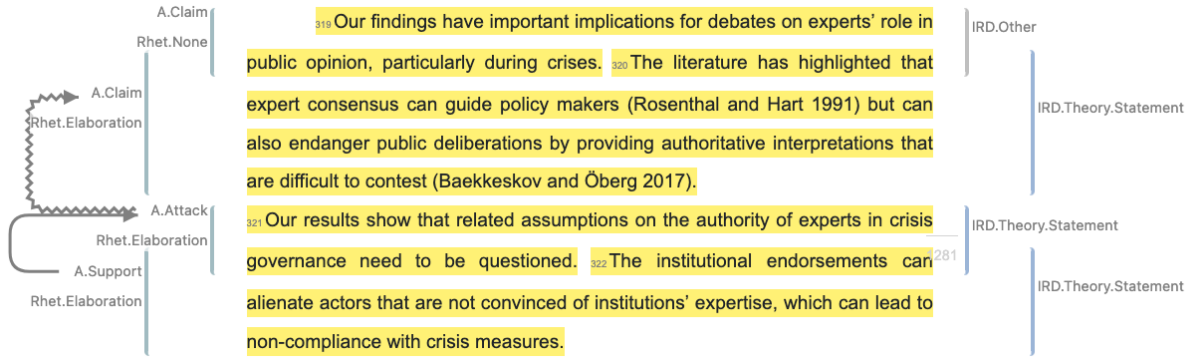
A *Definition* is a statement that explicitly specifies the meaning of a domain term.

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

Data statements provide empirical evidence, i.e. concrete references to the real world, for instance, to events (such as a war) or social facts.

A *Speculative* statement is about a possible present or future scenario or an alternative

(a) Example Paragraph with Complete Annotations



(b) Annotation Selector for a Markable

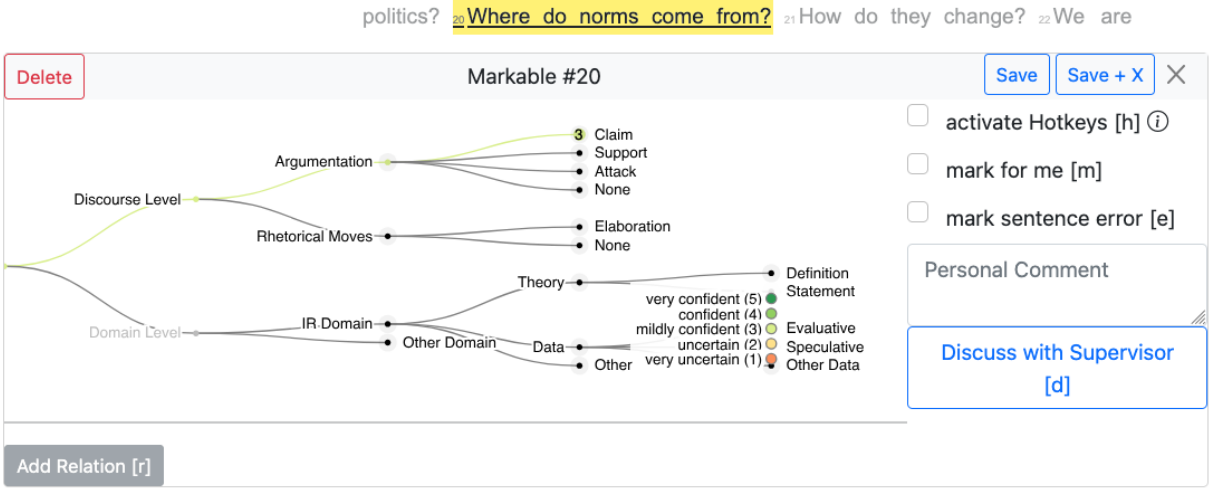


Figure 2: Annotation interface. **(a)** Fully annotated example paragraph. Argumentation and rhetoric dimensions on the right, domain discourse on the left. Straight arrows are support relations and zig-zagged are attack relations. **(b)** The annotation selector for each markable shows the class tree. Clicking a node sets the annotation and opens a confidence selector. Options to mark errors or questions are on the right.

past scenario, neither of which has actually happened.

An *Evaluative* statement references real-world events, data, or (social) facts and evaluates, interprets, or presents these theoretically.

Other Data are statements about any other real-world observations (e.g. *"Economic growth rates compound"*).

IR Other models IR-related statements that are neither data nor theory. This type is often used for statements about experimental design (e.g. *"In addition to OCHA officials, I interviewed individuals at the ICRC and UNHCR."*). Finally, *Other Domain* models statements that explicitly refer to a domain other than political science, IR, or global politics in general (e.g. *"The concept of heuristics stems from seminal research in cognitive psychology (Tversky and Kahneman 1974)"*).

4. Annotation Campaign

The SKILL-IR-Discourse corpus consists of two subsets, a Gold-standard set of five articles and a Silver-standard set of 185 articles, and was constructed in an extensive annotation campaign. The Gold-standard set was constructed first via a high-quality annotation of five selected articles, based on the methodology and annotation scheme from (Wolska et al., 2025). This set was used to train the annotators and develop the annotation tooling. The Silver-standard set is a selection of articles from all open-access publications (source articles) from major journals in international relations studies, balanced by five selection criteria and annotated by the trained annotators. A team of three political scientists consulted on the campaign.

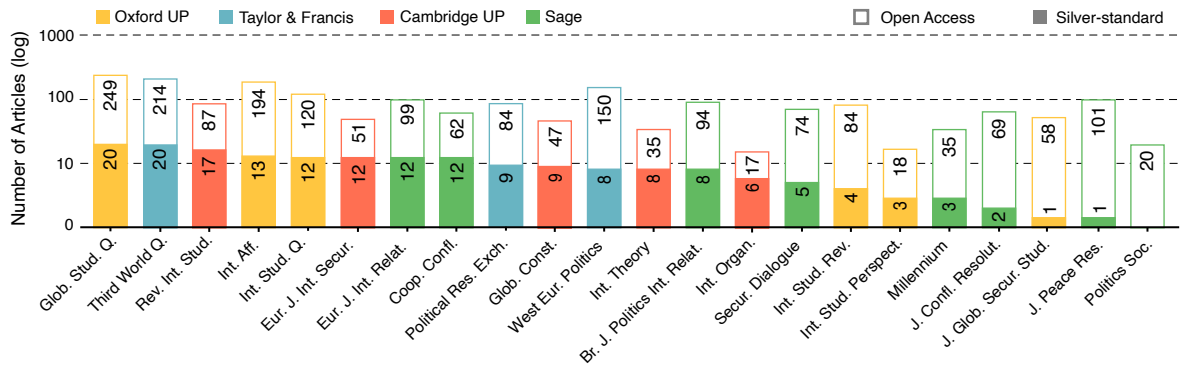


Figure 3: Number of articles per publisher in the source collection (gold open access) and included in the extended corpus (log scale).

4.1. Annotators

Two senior domain experts and student researchers carried out the annotation. The domain experts were responsible for annotating the Gold-standard corpus subset. The selection of trainee annotators was based on four criteria. First, their domain knowledge and understanding of IR theories, which was assessed in interviews and 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 of the hired annotators are studying in a political science program. Seven out of eleven annotators had a bachelor's degree in a relevant field (political science, politics and law, or cultural studies) at the start of employment. All of them were primarily educated in Western European universities. Four of the annotators were female and five were male. There was some fluctuation in the group of student annotators, but most of them have been with the project since the beginning. Currently, no discernible effect on annotation quality can be attributed to time of employment.

4.2. Annotation Tooling

We developed our own, specifically tailored annotation software. Existing tools such as Brat (brat.nlplab.org) or Label Studio (labelstud.io) did not meet the 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 and 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 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, our motivation was providing a visual appearance that mimics the look and feel of a typical scientific paper 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. Figure 2 shows (a) an annotated paragraph from a paper in our corpus and (b) a annotation label selector as they appear in the annotation tool.

4.3. Gold-standard and Annotator Training

As the Gold-standard subset for expert annotations and annotator training we selected five articles: four representative of one of the mainstream theories in the field of IR, namely Neo-Realism (Waltz, 1993), Liberalism (Putnam, 1988), Constructivism (Finnemore and Sikkink, 1998), and Feminism (Carpenter, 2005) as well as one paper representing a quantitative approach (Heinzel and Liese, 2021). The selection and focus on those articles was due to their relevance for the discipline as such and for the differentiation between the theories. The selected qualitative texts are considered central in the respective theoretical tradition. The articles were selected on the basis that they (1) introduce the basic building blocks of the theory, (2) make fundamental theoretical statements, and (3) are considered representative of their field. As these articles are fundamental texts and on the syllabi of many introductory courses on IR theories, we ensure the empirical relevance of the text selection for political science students. The quantitative text was included for the purpose of verifying applicability of the scheme and annotation methodology in general to a different text type

which is, however, also common in scholarly IR writing. All Gold-standard documents were manually segmented into sentences as markables using manual preprocessing as in (Wolska et al., 2025).

Gold-standard Annotations Annotation of the Gold-standard subset was performed with the annotation tool by two domain experts. Gold-standard annotations were obtained from the experts' annotations via disagreement mitigation: Cases of disagreement were discussed by domain experts and the project's linguist until a consensus was reached. Averaged 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 a prevalence problem.

Annotator Training and Quality Assurance The trainee annotators were provided with written annotation guidelines. They received systematic training in IR theories and category definitions and were supervised by domain experts and a linguist. Before annotating an article, the trainees were offered workshops on the background knowledge necessary to understand the theory represented in the article.

Each trainee annotator annotated all of the Gold-standard articles. Annotation quality assurance was performed as follows: Once all trainee annotators had finished annotating a given Gold-standard article, the true Gold-standard annotations were released, allowing the annotators to compare their annotations against them using the annotation tool's visual comparison mode. Then, annotators met in groups of two to three to discuss disagreements and prepare a list of questions about unclear annotations. These questions were sent to the supervisors. Supervisors met once a week to discuss the groups' questions and prepare explanations. Then, annotators and supervisors met once a week to clarify questions and discuss specific markables. Annotators received feedback on their performance in one-on-one meetings, during which patterns of deviations from the Gold-standard were pointed out. The average inter-annotator agreement for trainee annotators ranged from $\kappa = 0.36$ (fair) to $\kappa = .79$ (substantial).

4.4. Silver-standard Annotations

For the Silver-standard subset, annotated not by experts, but by trained annotators, we used texts selected from from source articles published by the four leading international publishers (Cambridge University Press, Oxford University Press, SAGE, and Taylor & Francis).

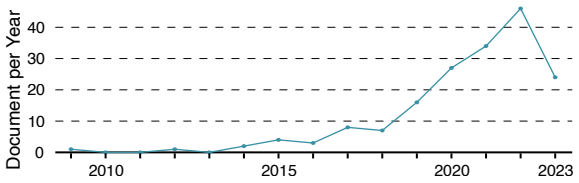


Figure 4: Number of documents included in the corpus split by year of publication.

Theory	
Critical Theory	33
Feminism	30
Liberalism	27
Constructivism	27
Realism	16
Institutionalism	13
Mixed	12
Neither	27

Author Gender	
Male	78
Female	67
Mixed	40

Method	
Qualitative	142
Quantitative	34
Mixed	9

Table 1: Corpus selection criteria with the number of articles matching each.

Source Article Retrieval As shown in Figure 3, we considered all articles that are published in one of the 22 journals from the four leading international publishers. Two publishers (Cambridge and Taylor & Francis) provided all respective articles on request as a collection of xml files in the JATS[1] standard. Articles from SAGE were collected in the same xml format from the publisher's API. Articles from Oxford were scraped from the publishers website in the html format using scrapy (`scrapy.org`).

All source articles were segmented into paragraphs using the original structure information (e.g. <p>). The paragraphs were then segmented into sentences, which served as the markable units (markables). We used Spacy's (`spacy.io`) sentenciser with the `en_core_web_sm` model, which we extended by several custom rules to correctly split various citation styles, multi-sentence inline quotes, and inline formatting information (e.g. <emph>). Annotators were given the option to mark remaining processing errors for manual correction.

Corpus Curation The Silver-standard subset was compiled out of high-quality scientific articles from the leading political science journals in field of IR mentioned above. Selected were only articles in English and those which are available for open access in order to maximize accessibility. Note that our subjective bias in the process of corpus selection reflects epistemic power asymmetries (bias) in various forms: for example, our choice of publishers reflects their dominant position in the field and the corpus is in a sense culturally tailored as

[1]NISO Journal Article Tag Suite, `jats.nlm.nih.gov`

it primarily represents Westernized contexts and English as the language. However, the choice of only open-access articles is an attempt to at least guarantee free access to the data.

The selected set was balanced as much as possible with respect to three factors: the applied research *method* (qualitative, quantitative, mixed), the *gender of the authors* (male, female, mixed authorship), and the core underlying IR theory (critical theory, feminism, liberalism, constructivism, realism, institutionalism, mixed, neither; where "neither" denotes articles not aligning with any one predominant theory). Each article in the Silver-standard subset was annotated by one trained annotator. The selection criteria along with distribution per category are shown in Table 1. Around 50% of the articles were single-authored and around 14% have more than two authors. Figure 4 shows the number of articles by year of publication, most of which are from recent years mainly because publishing as gold open-access is a relatively new development in political science.

Annotation Procedure for the Silver-standard Subset Silver-standard annotations stem from trained annotators extensively trained in the domain and the annotation categories and closely monitored during training as described in Section 4.3. Annotators were able to choose on their own which papers they would like to annotate from the curated corpus. Some annotators chose papers according to their domain interests after prereading them in the annotation tool and others proceeded with papers as they appeared in the list. Annotation quality throughout this phase of the project was assured as follows: Annotators were able to indicate and comment in the annotation tool that specific markables should be discussed with supervisors (see Figure 2(b)) and were explicitly encouraged to indicate markables for discussion as soon as they had any doubt in order to reduce annotation errors. Comments were reviewed by the experts, also in the annotation tool, and answered weekly. Throughout the annotation project, annotators were also able to communicate with the supervisors and the annotation interface creators via a Slack channel; in-person meetings were organized on as needed basis. Moreover, on a regular basis annotators met with a researcher on the project *not* involved in the corpus creation and the annotation process and were able to discuss non-domain topics both positive, such as learning and the benefits of participating in the annotation project, as well as negative, such as fatigue or frustration; the researcher communicated messages from the annotators to the annotation supervisors in an anonymized fashion ensuring annotators' comfort; proposals for improvements were implemented in as much as was possible.

	Gold-standard	Silver-standard
Documents	5	185
Paragraphs	396	9,559
Markables	2,491	54,605
Tokens	62,472	1,498,463
Claims	1,237	33,260
Relations	1,124	17,093
Support	0.95	0.98
Attack	0.05	0.02

Table 2: Corpus key figures.

5. Corpus Analysis

The SKILL-IR-Discourse corpus contains 190 articles, five of which are Gold-standard and 185 pf which are Silver-standard (see Figure 5). On average, the articles contain 52.4 paragraphs and 300.5 markables, each of which is densely annotated for argumentation, rhetorical moves and domain discourse.

We investigate two immediate analytical questions. Firstly, how frequently do the different types appear in political science articles, and what can we learn from their frequencies about the discourse of the field? Secondly, are the type frequencies stable across all documents in the corpus or do they vary? Can the variance only be attributed to the articles (i.e. it is a stylistic difference between authors), do external factors, such as the journal or method, influence the variance? Or is the variance an artefact of the annotation process?

5.1. Frequency Analysis

Our first question asks about the (differences in) usage frequency of the types. This can be measured via its frequency relative to the total number of markables. Figure 5(a) shows how these relative frequencies are distributed across all articles in the corpus. The following shows a comparison of the frequency of argumentative, rhetoric, and domain types (1) between documents (variance), (2) between types (mean difference), and (3) between corpus subsets.

Argumentation The most common argumentation types are *Claim* ($\mu = 0.49$ gold, 0.63 silver) and *Support* ($\mu = 0.43, 0.32$), while *Attack* ($\mu = 0.43$, 0.32) is exceedingly rare. This is to be expected in academic discourse, as the primary argumentative structure is a claim followed by supporting evidence. On average, fewer than half of the *Claim*s made are directly supported ($\mu = 0.38$; $\sigma = 0.07$ gold, $\mu = 0.28$; $\sigma = 0.12$ silver). About 5% of the

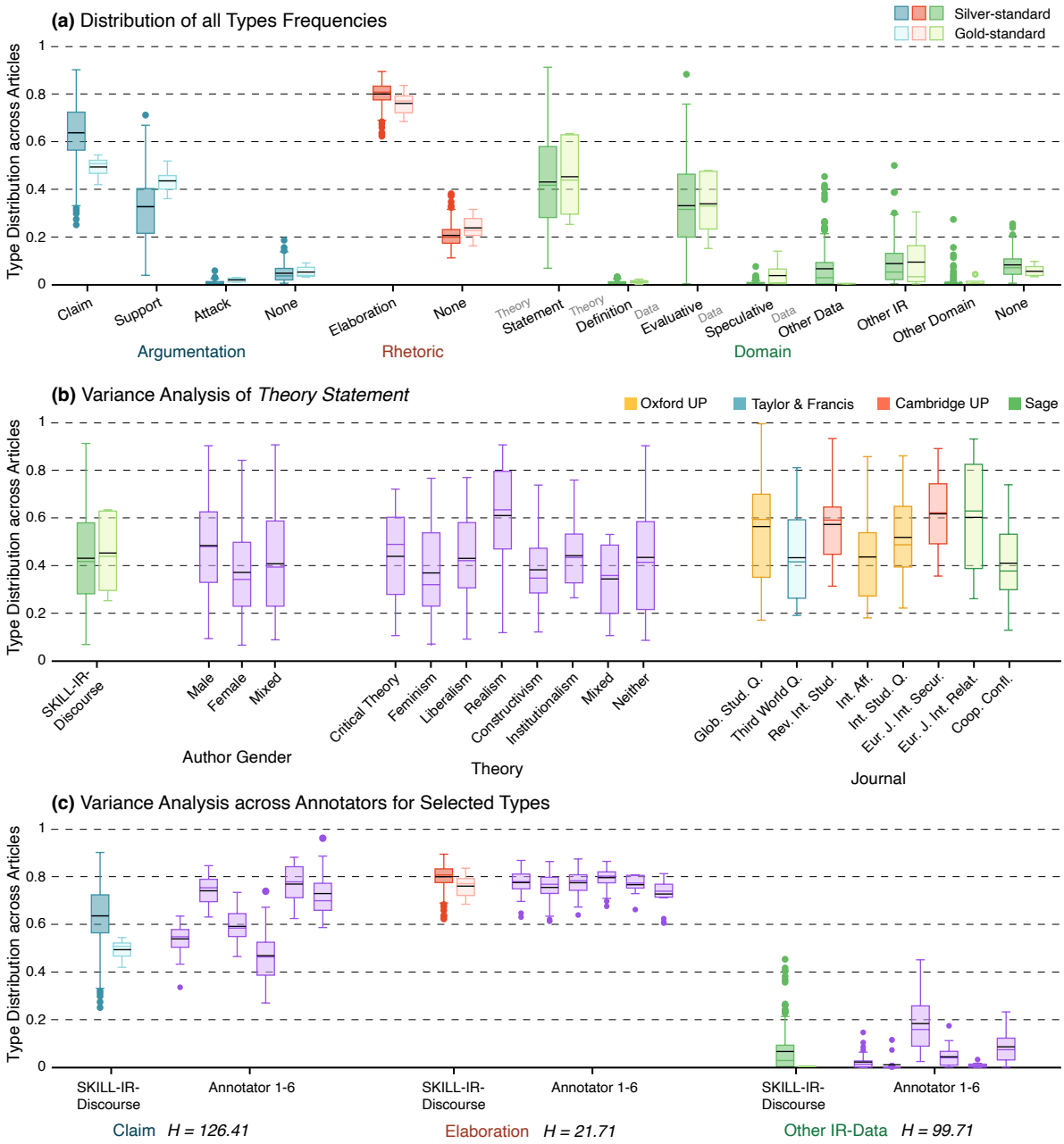


Figure 5: Distribution of the document-wide type frequencies. Each data point is the ratio of the respective type to *all markables* in one article. **(a)** Distribution of all types. **(b)** Distribution of the *Theory Statement* type across different author genders, theories, and journals. **(c)** Distribution of selected types across individual annotators compared to the complete corpus.

markables are not argumentative, such as structural information or academic jargon (e.g. "We examine our expectation in Table 1").

The main differences between the Gold- and Silver-standards are the frequencies of claims and supporting statements and the variance of these frequencies across documents. The Gold-standard has a lower variance, likely due to the more rigorous annotation. This suggests a degree of annotator bias in the Silver-standard. Additionally, the Gold-standard has a higher ration of *Support*s to *Claims*. This may be due to artefacts of the article selection process, given that Gold-standard articles are much older and have had a seminal impact. However, we cannot rule out the potential influence of the annotation procedure.

Elaborative structures Most markables across all documents ($\mu = 0.76, 0.80$) are minor claims that elaborate on some other (key/major) claim ($\mu = 0.19, 0.15$). This relationship remains consistent across documents and subsets. The schema includes only one rhetorical type, *Elaboration*, to distinguish between elaborative and key claims.

Element	Annotators		Journal		Method		Theory		Gender		Year	
	H	p	H	p	H	p	H	p	H	p	H	p
Claim	126.41	<0.001	11.04	0.137	1.59	0.662	6.85	0.445	4.53	0.104	1.45	0.918
Support	127.86	<0.001	12.18	0.095	2.35	0.504	4.04	0.775	5.41	0.067	1.87	0.867
Attack	60.93	<0.001	9.24	0.236	2.44	0.485	8.47	0.293	0.51	0.776	0.95	0.966
Elaboration	21.71	<0.001	13.13	0.069	0.69	0.875	3.34	0.852	5.43	0.066	9.54	0.089
Theory Statement	37.99	<0.001	15.97	0.025	1.76	0.623	15.38	0.031	11.23	0.004	2.25	0.814
Theory Definition	42.30	<0.001	20.30	0.005	4.43	0.218	17.00	0.017	1.17	0.556	3.49	0.625
Data Evaluative	24.12	<0.001	12.16	0.095	5.27	0.153	10.20	0.177	9.58	0.008	1.28	0.937
Data Speculative	21.50	<0.001	5.33	0.602	5.28	0.153	6.22	0.515	3.26	0.196	2.40	0.792
Data Other	99.71	<0.001	5.66	0.580	9.16	0.027	9.46	0.221	5.13	0.077	3.45	0.631
Other IR	75.37	<0.001	7.24	0.405	9.34	0.025	13.53	0.060	3.62	0.164	7.09	0.214
Other Domain	23.00	<0.001	15.0	0.036	5.37	0.147	14.03	0.051	4.38	0.112	6.48	0.263

Table 3: Effect sizes of the Kruskal-Wallis H-test across annotators for the H0 that all groups have the same mean, i.e. there is no difference. All elements are tested individually. Journals and Years with < 10 articles were excluded in the respective tests.

Variance across documents is low, as is the difference in mean and variance between the subsets.

Domain The *Theory Statement* ($\mu = 0.43, 0.45$; $\sigma = 0.16, 0.2$) and *Data Evaluative* types ($\mu = 0.33$; $\sigma = 0.13, 0.19$) are the foundation of the domain discourse across all articles. They have a high mean frequency and variance, and are consistent across the corpus subsets. Explicit term definitions (*Theory Definition*) and statements outside the field of political science occur sparsely ($\mu = 0.01, 0.03$) within the articles. Statements that are neither theory nor definition occur in some documents, but less frequently ($\mu = 0.11$; $\sigma = 0.09, 0.11$). *Data Speculative* ($\mu = 0.05, 0.01$) and *Data Other* ($\mu = 0.01, 0.09$) types are also infrequent within the documents. However, the mean frequency of both differs between the subsets: *Data Speculative* is more common in the Gold-standard set, while *Data Other* is more common in the Silver-standard set. As with the argumentation types, these differences may be attributed to a bias among the annotators or corpus subsets. The following variance analysis provides evidence for this hypothesis.

5.2. Variance Analysis of the Metadata

Our second question is whether there is a systematic explanation for the variance in the frequencies of the types, other than the author's style. This can be assessed via a variance analysis of the known metadata used to sample the corpus: eight journals (those with more than ten articles each), three methods, eight theories, three author genders, and five publication years (those with more than ten articles each). Table 3 shows the results of the Kruskal-Wallis H-test for all types and metadata with the H0 being that there is no difference between the groups. We use the H-test over ANOVA because the variances differ (see Figure 5(b)).

For all argumentative and rhetoric types $p >> 0.05$, in most cases strongly so. Therefore, we can assume that there is no difference between the groups for these types. For domain types, however, H0 can be rejected in several cases. The most common type, *Theory Statement*, shows significant differences between journals, theories, and author gender with $p < 0.05$. Figure 5(b) shows the distribution of *Theory Statement* for each class in these three groups. The second most common type, *Data Evaluative*, shows significant differences between author genders and, with a lower significance threshold ($p < 0.1$), also between journals. Different journals in the field are known to prefer different kinds of work, e.g. empirical, theoretical, or methodological. It is a valid assumption that the discourse style differs accordingly. The differences between author gender and theory are more difficult to explain without more detailed study.

It should be noted that the tests have not been corrected for multiple testing, so there is a small risk that the significance is accidental. For this reason, we refrain from analyzing the variance of the very rare types.

5.3. Variance Analysis of the Annotators

We also analyze the variance between different annotators and find that they explain a significant part of the total variance of the types. Table 3 shows that there is a significant difference between the annotators for all types ($p << 0.001$). The effect sizes are particularly large for the argumentation types, which explains why the variance of there is much greater in the Silver- than in the Gold-standard. Figure 5(c) illustrates the distributions for (*Claim*, *Elaboration*, and *Other IR Data*).

The distribution for *Claim* shows that annotators have strongly different preferences when deciding between claims and supports with very high effect

sizes. However, each annotator is also consistent and the variance of each annotator is comparable to that of the Gold-standard. The remaining variance between annotators of about 0.2 is unexplained, possibly due to the subjectivity of discourse perception.

The distribution for *Elaboration* differs significantly between the annotators, but the effect size is lower. Consequently, the means and variances are more similar than those for claims, resulting in low overall variance of elaborations.

The distribution of *Other IR Data* is also significantly different. While most annotators use these rare types very sparingly, others use them much more frequently. This leads to some types (*Other IR Data, Other Domain, Other Data, and Data Speculative*) having a low mean frequency, but many outliers.

It should be noted that the annotator variance may contain selection biases, as all annotators were asked to select the articles they wanted to annotate.

6. Conclusion

In this work, we present a large annotated corpus comprising 190 political science articles with dense, sentence-level annotations concerning argumentation and domain-specific discourse. We describe our approach to curating the corpus, the annotation campaign and the tools used, as well as our method for training and evaluating annotators on a challenging annotation task. Finally, we analyze the corpus to gain insight into domain-specific argumentation and discourse, and to understand the reasons behind variations between articles.

We find that our articles mostly consist of claim-support structures, in which fewer than half of all claims are supported. Most claims are minor and elaborative in nature, and attack relations are exceedingly rare. The claim-to-support ratio varies widely between articles, due to a combination of annotator bias and authorial style. No significant influence of external factors on the argumentative structure was found.

The domain discourse largely consists of theory statements and data evaluation, the frequency of which varies strongly between articles. As before, this variance can be partially attributed to annotator bias. However, author style and external factors play a larger role than argumentation types. Some types occur significantly more or less frequently depending on external factors such as the publishing journal, method, theory and, in some cases, author gender.

7. Bibliographical 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.

R. Charli Carpenter. 2005. "women, children and other vulnerable groups": Gender, strategic frames and the protection of civilians as a transnational issue. *International Studies Quarterly*, 49(2).

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

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

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.

Mirko Heinzel and Andrea Liese. 2021. Expert authority and support for covid-19 measures in germany and the uk: a survey experiment. *West European Politics*, 44(5-6).

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*.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, Brussels, Belgium. ACL.

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.

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.

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

François Randour, Julien Perrez, and Min Reuchamps. 2020. Twenty years of research on political discourse: A systematic review and directions for future research. *Discourse & Society*, 31(4).

Maria Sidiropoulou. 2013. Representation through translation: Shared maps of pragmatic meaning and the constructionist paradigm. *Journal of pragmatics*, 53.

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*.

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.

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.

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

Magdalena Wolska, Bernd Fröhlich, Katrin Girgensohn, Sassan Gholiagha, Dora Kiesel, Jürgen Neyer, Patrick Riehmann, Mitja Sienknecht, and Benno Stein. 2024. Argumentation in Waltz's 'Emerging Structure of International Politics'. In *10th Conference of the International Society for the Study of Argumentation (ISSA 2023)*,

pages 969–979, Amsterdam. Sciential International Centre for Scholarship in Argumentation Theory.

Magdalena Wolska, Sassan Gholiagha, Mitja Sienknecht, Dora Kiesel, Irene Lopez Garcia, Patrick Riehmann, Matti Wiegmann, Bernd Fröhlich, Katrin Girgensohn, Jürgen Neyer, and Benno Stein. 2025. Argumentation and Domain Discourse in Scholarly Articles on the Theory of International Relations. In *31st International Conference on Computational Linguistics (COLING 2025)*, pages 9238–9249. Association for Computational Linguistics.