Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Un cadre d'annotation pour la modelisation logique de l'argumentation politique

Cecilia Graiff[1]
(1) ALMAnaCH, Inria Paris
cecilia.graiff@inria.fr

RÉSUMÉ

Cet article présente un cadre destiné à la construction en cours d'un corpus multilingue annoté selon sa structure argumentative. Il s'inscrit dans un projet de thèse visant à modéliser les schémas argumentatifs dans des données politiques à l'aide des (grands) modèles de langue. Cette recherche poursuit deux objectifs principaux : (1) étudier l'argumentation en contexte multilingue, en prenant en compte les variations liées à la langue, au pays et à la culture ; et (2) établir un pont entre l'argumentation fondée sur la logique et le traitement automatique des langues, en intégrant des mécanismes de raisonnement logique dans la chaîne de traitement afin d'améliorer la généralisabilité des modèles. Un défi majeur réside dans le manque de jeux de données à la fois multilingues et annotés selon la structure logique des arguments. Pour y remédier, nous avons collecté des données multilingues et publié des expérimentations dans un cadre bilingue. Cet article introduit une stratégie d'annotation fondée sur la logique, basée sur le cadre d'argumentation ASPIC+ (Modgil & Prakken, 2014), ainsi que des expérimentations préliminaires utilisant des modèles de type Transformers.

ABSTRACT

An annotation framework for the logical modelling of political argumentation

This paper presents a framework intended for the ongoing construction of a multilingual corpus annotated according to its argumentative structure. It is part of a Ph.D. project that models argumentation patterns in political data using (large) language models. The research pursues two objectives: (1) to investigate argumentation in multilingual settings, accounting for differences in language, country, and culture; and (2) to bridge logic-based argumentation and natural language processing by integrating logical reasoning into the pipeline to improve generalizability. A major challenge is the lack of datasets that are multilingual and annotated according to the logical structure of arguments. To address this, we collected multilingual data and published experiments in a bilingual setting. This paper introduces a logic-based annotation strategy based on ASPIC+ argumentation framework (Modgil & Prakken, 2014) and preliminary experiments with Transformers-based models.

MOTS-CLÉS : Modèles de langue, modélisation logique de l'argumentation, transfert interlinguistique, argumentation politique.

KEYWORDS: Language models, logical modelling of argumentation, cross-lingual transfer, political argumentation.

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 186–201.

1 Introduction

The analysis of argumentation structures has been a topic of research in natural language processing for many years. However, the predominant approach has always been based on linguistic cues (Lawrence & Reed, 2020). This included teaching the model recurring patterns that usually mark argumentative sentences, such as the presence of connectives like “hence“ or “therefore“. In more recent times, with the advent of large language models (LLMs), the reasoning structure of argumentative texts started to be the main research focus (Li *et al.*, 2025). However, the research in this field is still very limited to monolingual settings, with only a few works focusing on multilinguality; even then, most of these works leverage an automatic translation pipeline, thus demonstrating the need for multilingual resources.

Therefore, we present here the first steps of a thesis that aims to (1) investigate the performance of argument detection strategies in multilingual and cross-country settings, and (2) integrate a logic-based argumentation framework to ground our research in the logical structure of arguments, rather than relying solely on linguistic cues. Our hypothesis is that these two research directions can come together in paving the road to more robust and generalizable argument mining algorithms, not biased by changes in language and culture.

We focus on the political domain, as it allows a comparative analysis of language- and country-related differences. The preliminary experiments presented in this paper are bilingual (English and French), and we aim at extending them to cover other languages, specifically Spanish, German, and Italian. However, this work poses the significant challenge of the absence of annotated datasets in the domain: to the best of our knowledge, the only available dataset is ElecDeb60to20 (Goffredo *et al.*, 2023), which only includes US English data and is not annotated specifically for argument schemes. We argue that, in order to be able to take into account logical cues when modeling and extracting argumentation structures with language models, an annotation of argument types is necessary. For this reason, we present in this paper the annotation framework behind FrenchPolArg, a new corpus of political discourse in French annotated following the logical structure of its arguments.[1] The proposed annotation framework aims at filling a research gap, as datasets are either annotated from a linguistic point of view, or not focusing on the political domain. Moreover, resources in this field are predominantly available in English, which highlights the innovation of a French dataset.

In this paper, we present the following contributions:

1. An annotation framework derived from the ASPIC+ argumentation framework (Modgil & Prakken, 2014) and applied to natural language arguments in order to enable the mapping of reasoning patterns of argumentative texts with (large) language models based on logical rules.
2. An annotation strategy for building FrenchPolArg, a French political annotated dataset, including retrieval pipeline, annotation protocol, and planned quality control.
3. We openly release a seed dataset and provide pilot evaluation of a first semi-automatic annotation process for argument components. This part was more extensively described in (Graiff *et al.*, 2026).

The paper is structured as follows: we first review the state of the art in multilingual argument mining

[1] A preliminary version of the corpus is released on GitHub under the MIT licence. We plan on releasing the whole dataset after the end of the annotation campaign.

and logic-based argumentation modeling. Next, we discuss argumentation datasets, highlighting the lack of political resources. We then present our annotation framework, formalizing natural language arguments, and describe the creation of a new French political dataset. We briefly report a transfer learning experiment examining performance drops across language and cultural shifts, which laid the foundations for a semi-automatic annotation workflow (Graiff *et al.*, 2026). Finally, we outline our future plans to use this resource as ground truth to test the integration of logic-based methods into argumentation mining workflows in a cross-lingual and cross-country setting.

2 Related Work

Traditionally, argument mining has been modeled as a pipeline of subtasks (Li *et al.*, 2025), including argumentative segment detection, component classification (e.g., claims and premises), and relation identification, as outlined by Stab & Gurevych (2017). This section reviews prior work most relevant to this thesis, focusing on multilingual argument mining, logic-based modeling, and defeasible reasoning in large language models, before concluding with an overview of existing datasets to highlight the need for a new resource.

Cross-lingual argument mining (Eger *et al.*, 2018) applied translation and annotation projection to benchmark the cross-lingual performance of language models. (Toledo-Ronen *et al.*, 2020) also employ translation in their pipeline for stance detection alongside other argument mining-related tasks such as argument quality. Interestingly, they show that translated English data can even improve performance on English itself. Similarly, (Yeginbergen *et al.*, 2024) demonstrate that in the case of multilingual argument mining, data-transfer methods outperform model-transfer. Their study is a further proof of the tendency of language models to learn data rather than the argument mining task itself. (Schaefer *et al.*, 2022) experiment with claim detection with BERT and RoBERTa on several dataset combinations in order to find the best composition for training, which appears to have large corpus size, homogeneous claim proportions, and less formal text domains. (Ruiz-Dolz *et al.*, 2024) leverage transfer learning to investigate the argument relation identification task in cross-lingual and cross-domain settings. While showing an improvement, their results also highlight the problem of cross-lingual robustness. Differently from the previously cited works as well as our own, they rely on manually annotated dataset, and do not integrate experiments with automatic translation and annotation projection. From these works, it is evident that the multilingual aspect is only taken into account in a minority of the papers in the field, mostly due to the lack of resources and to the difficulty of models to generalize the task. We therefore aim at filling this research gap by first providing a suitable resource, and then delivering a logic-based multilingual approach.

Logic-based argumentation modeling The works that approach the study of argumentation in natural language texts from the perspective of their logical structure are rare. Most argument mining research focused on training models to learn linguistic cues such as the connectives that tend to be present in claims or premises (Lawrence & Reed, 2020). However, since the advent of Large Language Models (LLMs) more attention has been given to detecting the reasoning pattern, for example by leveraging Chain of Thought or other reasoning strategies. (Lalwani *et al.*, 2025) leverage Llama-7B, GPT-4o, GPT-4o-mini and OpenAI o1-preview to autoformalize natural language to first order logic, and later validate the arguments with Satisfiability Modulo Theory (SMT) solvers. (Lei & Huang, 2024) leverage the linguistic connectives present in the ten most common logical structures of argumentative reasoning to build a logical tree, and implement LLMs to translate the tree into natural

language, with the final aim of detecting logical fallacies. The translation into a natural language description, together with the tree's embeddings, is further used to perform fallacy detection and classification. This approach shows improvements over traditional fallacy detection methods, with the F1 score for fallacy detection reaching 87.19 for Llama-2 on the best-performing dataset, and fallacy classification reaching 83.95. (Helwe *et al.*, 2024) approach the closely related fallacy detection task from the point of view of logical structure, by building a taxonomy of fallacies and then evaluating several language models under a zero-shot learning setting. The best results are reported by GPT 3.5 175B, which has a F1 scores of 0.627 in the binary classification setting; however, this results is still way below the manual score of 0.749 reported by averaging human annotations of samples. This further motivates the need for a new annotated resource to test logic-based modeling of real-world arguments.

Defeasible reasoning and LLMs The main challenge of applying logical reasoning to real-world data from the political domain is the scarce correspondance between the abstract framework and the possibly faulty or ill-formulated natural language reasoning. For this reason, we rely on defeasible logic, a non-monotonic logic used to formalize defeasible reasoning (Nute, 1994) which is characterized by the possibility of retracting or changing claims based on lately acquired knowledge, as well as allowing for exceptions, context-dependent rules, and changes, differently from first-order or propositional logic. We chose to rely on this formalization because it allows modeling imperfect real-world arguments. Defeasible reasoning has been analyzed in LLM research independently from argument mining. (Allaway & McKeown, 2025) present the DEFREASING dataset to evaluate defeasible reasoning about property inheritance and evaluate 12 instruction-tuned LLMs, with a best F1 score of 0.64. We note that differently from them, we aim at building a real-world dataset, while their example mostly mimick basic logical patterns such as syllogisms. (Tachmazidis *et al.*, 2024) benchmark various defeasible rule-based reasoning patterns by translating defeasible rules into text suitable for LLMs. They focus on a qualitative analysis and conclude that while the results are encouraging, especially with GPT-4o, there is still a lot of work to be done. (Fang *et al.*, 2025) provide an interesting work that brings together LLMs and argumentation frameworks, namely the ASPIC+ framework (Modgil & Prakken, 2014) that we also leverage in this project. They use LLMs to parse raw textual inputs into structured elements (Facts, Rules, and Preferences), obtaining the best results with Gemini-2. We aim at bridging these methods with argumentation mining, thus evaluating defeasible reasoning approach to argumentative reasoning.

Datasets One of the most widely used argument mining datasets is the UKP Sentential Argument Mining corpus (Stab *et al.*, 2018), which consists of more than 25,000 sentences from heterogeneous texts over eight topics with annotations done by crowdworkers. Each sentence is annotated as argumentative or non-argumentative, and the stance is included. Contrarily to our aim, this dataset does not provide inferential types. Its heterogeneous nature is also not suited to our aim, as it merges different domains and registers. Another example is AbstRCT (Mayer *et al.*, 2021), which focuses on the medical domain and reports annotations for argumentative components and relations. The annotations of the Araucaria datasets (Reed & Rowe, 2004) are also based on argument schemes, but the dataset is neither multilingual nor from the political domain.

Defeasible reasoning approaches often draw upon datasets specifically developed to support them: for example, (Fang *et al.*, 2025) use BoardGameQA (Kazemi *et al.*, 2023), a synthetic text-based logical reasoning dataset, and introduce MineQA, a synthetic dataset created to extend the methodology to a more comprehensive evaluation of defeasible reasoning in natural language. Differently from these examples, we aim at analyzing real-world data from the political domain, which involves the challenge of finding an accurate correspondance between the abstract rules of classical and defeasible

logic, and the concreteness of - oftentimes faulty - real life reasoning.

To the best of our knowledge, the most important existing annotated resource in the political domain is ElecDeb60to20 (Haddadan *et al.*, 2019; Goffredo *et al.*, 2023), a collection of US presidential debates from 1960 until 2020 scraped from the website of the Commission on Presidential Debates [2]. A first version was released in 2016 with annotations of claims, premises, and relations, and in 2020 a later version including fallacy annotation was published. These last kind of annotations are closer to our aim of detecting inference type; however, the dataset focuses on fallacious patterns rather than general argumentation structures. While our work leverages this contribution, we investigate two different topics: multilinguality and logic-based argumentation. A notable contribution is also FredSUM (Rennard *et al.*, 2023), which partially overlaps with the dataset that we present. However, because FredSUM was created with the aim of experimenting with summarization, it is divided into topics, which is not suitable for our aim; moreover, it is not annotated for argument structure. Hence, we affirm that due to the intersectionality of our research, spanning between natural language processing, classic and defeasible logic, and an application to real-world political data, a new resource is necessary to fill the gap in logic-based argumentation modeling.

3 Annotation framework

This section outlines the conceptual foundation of our work. We envisage a logic-based extraction of argumentative reasoning patterns with LLMs, hence our decision to follow formal argumentation models. The main challenge for the annotation task lies in bridging formal models and real-world reasoning structures, which often do not follow strict logical rules. We start by presenting ASPIC+, the formal argumentation framework to which our work is oriented, in Section 3.1. In Section 3.2, we present the argument formalization at the core of the annotation pipeline, based on the formal model presented in 3.1, and adapted to real-world political data.

3.1 Theoretical framework

Argumentation frameworks are formal models that represent arguments and their conflicting relationships. Due to its leverage of defeasible rules, we orient our definitions to the ASPIC+ framework, and recall its main concepts in this section, following (Modgil & Prakken, 2014).

Definition 3.1. *An argumentation framework is a triple* $\mathcal{A} = (\mathcal{L}, \mathcal{R}, n)$*, where:*

- $\mathcal{L}$ *is a logical language closed under negation* ($\neg$)*.*
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ *is a set of strict* ($\mathcal{R}_s$) *and defeasible* ($\mathcal{R}_d$) *inference rules.*
- n *is a partial function such that* $n : \mathcal{R}_d \to \mathcal{L}$*.*

We also introduce the concept of a knowledge base, which contains strict and defeasible rules.

Definition 3.2 (Knowledge Base)**.** *Given an argumentation system* AS*, a* knowledge base K *is a finite set of formulas from* L *consisting of:*

[2] https://www.debates.org/voter-education/debate-transcripts/

- *A set K_n of* axioms *(undisputed facts);*
- *A set K_a of* ordinary premises *(assumptions that can be attacked).*

For a given argument, the function $Prem$ returns all the formulas of K (called premises) used to build the argument, $Conc$ returns its claims[3]. Arguments are defined as follows:

Definition 3.3 (Argument)**.** *Given an argumentation system AS and a knowledge base K, an* argument *A is a finite tree with nodes labeled by formulas from L, defined recursively:*

- *Any $p \in K$ is an argument with* $\mathrm{Prem}(A) = \{p\}$ *and* $\mathrm{Conc}(A) = p$*;*
- *If $A_1, \ldots, A_n$ are arguments and $r \in R_s \cup R_d$ is a rule with conclusion q such that the premises of r match the conclusions of $A_1, \ldots, A_n$, then there is an argument A with* $\mathrm{Prem}(A) = \bigcup_i \mathrm{Prem}(A_i)$ *and* $\mathrm{Conc}(A) = q$*.*

We note that this definition of arguments is particularly suitable to our research, as it allows modeling imperfect reasoning structures such as premises that overlap with claims, or claims that support other claims (modeled as subarguments). As we aim at modeling the argumentation structure, we are also interested in the way arguments interact with each other. Therefore, we report the attack relations, classified in undermining (attacking premises), rebutting (attacking claims), and undercutting (attacking inference rules):

Definition 3.4 (Attack)**.** *A attacks B iff A undercuts, rebuts, or undermines B, where:*
Undercuts: *A undercuts argument B (on B') iff $Conc(A) = -n(r)$ for some $B' \in Sub(B)$ such that B''s top rule r is defeasible.*
Rebuts: *A rebuts argument B (on B') iff $Conc(A) = -\varphi$ for some $B' \in Sub(B)$ of the form $B'_1, \ldots, B'_n \Rightarrow \varphi$.*
Undermines: *A undermines B (on ϕ) iff $Conc(A) = -\varphi$ for an ordinary premise ϕ of B.*

Based on ASPIC+, we consider real-world arguments composed of premises and claims. Strict and defeasible rules model the connections among premises and claims, with the main feature of allowing faulty reasoning. Lastly, attack relations among the arguments define the argumentation structure.

3.2 Argument formalization

This section describes the planned pipeline of this thesis project, based on an adaption of a formal argumentation framework to natural language text. We highlight that our purpose is not to fully replicate ASPIC+, but rather to use it as a reference for logical-based argumentation research in multilingual settings. Therefore, we do not exclude expanding the presented pipeline with additional features derived from ASPIC+, such as preferences. Following the concepts presented in section 3.1, we define three different levels of the extraction task.

Level 1: Premises and Claims. Both in formal argumentation and in argument mining, arguments are considered to be composed of premises and claims, as explained in Definition 3.3. Defined

[3]Traditionally, formal argumentation frameworks such as ASPIC+ and Walton's schemes leverage the term "conclusion", which for clarity's sake we replace here with "claim", the standard term in argument mining, and the one term used throughout this paper.

as argument components, these are the building blocks of arguments, and their extraction is a necessary first step. We define this as a labeling task, where $\mathcal{S}_d$ is the set of all spans of a given document d, and a span is an index pair $[i, j]$ with $1 \leq i \leq j \leq |d|$. We define a labeling function $\ell_d : \mathcal{S}_d \to \{\text{Claim}, \text{Premise}, \text{Not Argumentative}\}$.

While this is the part where linguistic connectors are the most useful to the detection process, we aim at a precise annotation, which requires understanding of the logical structure of the text. For this reason, we highlight the importance of (1) involving human annotators, and (2) following a formal framework such as ASPIC+. We report an example of annotated claims and premises in Figure 1.

We further assume that the axioms and assumptions described in the ASPIC+ Knowledge Base (Definition 3.2) are partly explicit and partly implicit. For this reason, further experiments will integrate this step with an extraction of common knowledge facts from cultural datasets; however, we do not discuss this further in this paper, as it does not affect the annotation framework.

D'abord [**je veux vous dire que ce que vous avez dit est factuellement faux**]. [*Le bouclier que nous avons mis en place, les chiffres sont là pour le constater fait que la France a deux fois moins d'inflation que l'Espagne, 60% de moins que l'Allemagne et qu'à peu près tous nos voisins*]. [***Parce que le bouclier qu'on a mis en place, au contraire il n'est pas inflationniste***]. [*Il évite que la hausse que l'on ne répercute pas sur les ménages se traduisent par ailleurs*]. [**Donc ce que vous avez dit, c'est l'exact contraire des faits vérifiables aujourd'hui**].

Figure 1: Example of claims and premises from the 2022 Macron–Le Pen debate. Claims are in **bold**, premises in *italics*, and all argumentative segments are in square brackets. One segment serves as both claim and premise.

Level 2: Rules. In our application of the ASPIC+ framework, the strict and defeasible rules described in Definition 3.1 correspond to the second level. They are based on the support relations among the premises and claims, and we define them with $\mathcal{R}$.

We rely on (Walton *et al.*, 2012) and (Walton & Hansen, 2013) to define 15 argument schemes to be annotated, reported in Appendix A. Argument schemes provide the logical rules behind different types of arguments, and as such we use them to represent the rules defined by (Modgil & Prakken, 2014). The schemes were chosen based on their frequence in natural language texts and the possibility of formalizing their logical structure. (Walton & Hansen, 2013) is often considered a faithful formalization of natural language arguments due to its use of defeasible logic, which we consider suitable to model the imperfect reasoning structures that are typical of real-world arguments. We focus on extracting the logical structure of arguments, not their validity. Some schemes, like *ad hominem*, are considered fallacies elsewhere, but here we only trace the logical chain of premises and claims. While not relevant to this paper, a validation step may be added later.

As an example, we focus on the excerpt from the Macron-Le Pen debate reported in Figure 1. This is an example of argument from sign, because the speaker arguments based on empirical evidence, namely the reported statistics for inflation in European countries.

Level 3: Attacks. On the third level, we investigate the attack relations among arguments, classified into undermining, rebutting and undercutting as described in Definition 3.4. We thus define attacks as

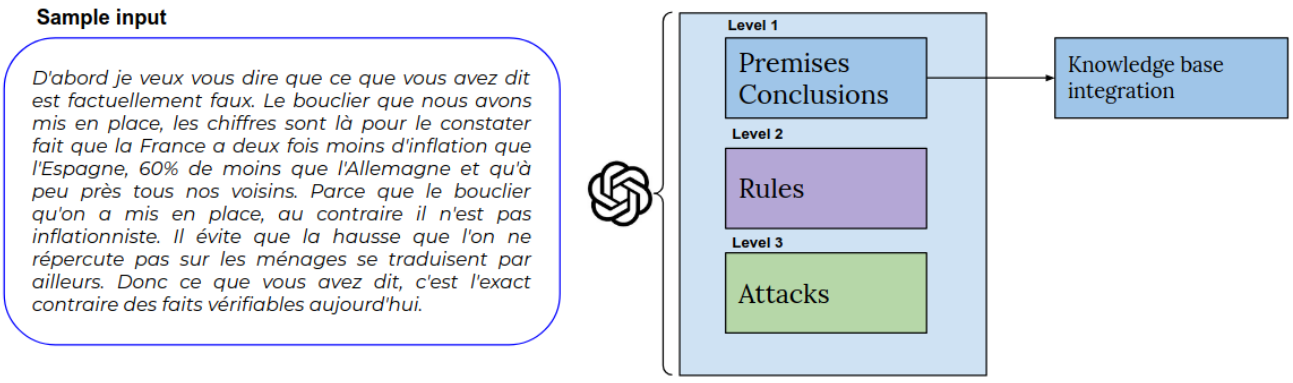

Figure 2: The planned workflow for which the dataset will be leveraged.

follows:

$$\text{Attack}(d) = \{(A_i \mid A_i \in \mathcal{AS})\} \times \{(A_j \mid A_j \in \mathcal{AS})\} \times \{undermine, rebut, undercut\},$$

where $\mathcal{AS} = (\ell_k, \mathcal{R}_k, Att_k)$ is the set of argument components, rules, and attack relations.

We aim at annotating and extracting undermining (attacking premises), rebutting (attacking claims), and undercutting (attacking inference rules), with the purpose of building argument graphs and eventually performing a comparative analysis between argumentation styles across languages and cultures. Political debates constitute a particularly suitable dataset, as the participants tendentially aim at demolishing each other's arguments.

A visualization of the three levels is presented in Figure 2.

Multi-level dataset definition Let $\mathcal{D}_{gold}$ be a natural language corpus annotated following the ASPIC+ framework, composed of documents $d_1, ..., d_n$. To perform the three steps of the pipeline, we need to define $\mathcal{D}_{gold}$ as a multi-level corpus, marked as gold standard with a star ($\star$):

$$\mathcal{D}_{gold} = \{(d_k,\ y_k^{(1)},\ y_k^{(2)},\ y_k^{(3)})\}_{k=1}^{n}$$

where each level carries the gold output for the corresponding subtask:

$$y_k^{(1)} = (\mathcal{S}_k^\star, \ell_k^\star) \quad \text{— gold spans and labels}$$

$$y_k^{(2)} = R_k^\star \quad \text{— gold rules}$$

$$y_k^{(3)} = \text{Att}_k^\star \quad \text{— gold attacks}$$

Therefore, the implementation of the planned approach requires annotating a dataset for ground truth, following the three levels described in this section. In the next section, we report the annotation schema followed to annotate the three levels of the gold standard corpus.

4 Data collection and annotation protocol

This paper aims at testing a logic-based pipeline for argumentative reasoning on multilingual, cross-country data. Since available resources exist only in English, we retrieved additional data, starting with bilingual French–English experiments. We outline in this section the retrieval pipeline, annotation tasks, and planned quality-control strategies.

Retrieval pipeline We published FrenchPolArg, an annotated French corpus of political debates and speeches. FrenchPolArg was collected by scraping the transcripts of the Presidential debates from 1974 until 2022 available on Vie Publique. In the cases where a transcript did not already exist, the corresponding YouTube video was scraped and diarized with WhisperX, which provides fast automatic speech recognition (70x realtime with large-v2) with word-level timestamps and speaker diarization (Bain *et al.*, 2023). In addition to the election debates, we also scraped the presidential speeches available on the same website in form of PDF documents. Due to the noisy nature of the files, a thorough cleaning task was required. A preliminary and partially annotated version of FrenchPolArg is freely available on GitHub.

Dataset FrenchPolArg consists of 8 presidential debates and 21 presidential speeches. Each debate is animated by 2 journalists and 2 candidates to the presidential elections.

Statistic	Value
Number of debates	8
Number of speeches	21
Number of debate speakers (total)	23
Number of tokens (debates)	202,046
Number of tokens (speeches)	88,885

Table 1: Summary of dataset statistics for the debates collected in FrenchPolArg. Because the preliminary experiments described in section 5 focused on the debates, we report here the statistics of this portion of the dataset.

Tasks We divided the annotation process in several subtasks, following the level-based schema of the planned approach (Figure 2). Firstly, we focused on detecting argument components, as we consider them the primary block of argumentation structures. Based on the formalization expressed in section 3.2, we launched an annotation campaign by recruiting expert annotators at SciencesPo Paris. We provided them with annotation guidelines and are currently running this campaign, which should soon deliver the expected results. Secondly, our annotation regards the annotation schemes reported in Table 3. Lastly, we want to annotate the attack relations among arguments, to be able to build an argumentation graph.

Annotation process We are currently hiring two annotators, who will annotate the same data in order to provide an annotation agreement. They use the INCEpTION platform (Klie *et al.*, 2018) and

perform token-level annotations, following the Beginning-Inside-Outside (BIO) schema (Ramshaw & Marcus, 1995). The annotators are provided with annotation guidelines that follow the framework presented in this article, and deliver specific guidance by asking critical questions. The first batch of annotations (composed by three debates that were already annotated by the author of this paper) is used as training for making sure that the annotators understand the task, and delivering preliminary results on the annotation agreement, in order to eventually adjust the guidelines.

Planned quality control Each sequence is annotated by two expert annotators, and Krippendorph's Alpha will be used as a metric to measure inter-annotators agreement. In case of disagreement, the annotation will be compared to the one obtained with language models. We deliver preliminary insights into a semi-automatic annotation workflow in (Graiff *et al.*, 2026), and will leverage language models according to our findings. In case of complete disagreement even with the integration of language models, the author of this paper will analyze the sequences and choose the relevant label.

5 Experiments on the Component Detection Task

We present here the preliminary experiments performed on the component detection task, corresponding to Level 1 of the pipeline shown in Figure 2. We propose a semiautomatic extraction approach by leveraging transfer learning techniques with masked language models, with the aim of finding the best-performing approach for integrating language models in the annotation process. We automatically translated ElecDeb60to20 (Goffredo *et al.*, 2023) to French with Opus-MT (Tiedemann & Thottingal, 2020), and described the annotation projection process in (Graiff *et al.*, 2026), where we also benchmarked both the model transfer setting (leveraging the cross-lingual capabilities of the model itself) and the data transfer setting (leveraging the translated resource to avoid the language shift effect).

To enable pilot experiments, the author of this paper performed a preliminary annotation of 3 debates from FrenchPolArg, for a total of 68,392 tokens. The annotated debates were chosen with the aim of minimizing possible cross-lingual biases: for this reason, we annotated the first (Giscard d'Estaing - Mitterrand, 1974), middle (Chirac - Jospin, 1995), and last (Macron - Le Pen, 2022) presidential debate. An evaluation of the first annotation batch is present in Table 2.

Test dataset	Macro F1	Micro F1	F1 Premise	F1 Claim
FR-ElecDeb60to20	0.55	0.56	0.51	0.46
FrenchPolArg	0.48	0.50	0.44	0.37

Table 2: Results of argument component detection with mBERT, trained on EN-ElecDeb60to20.

Our results confirm a performance drop under language shift. We report a macro F1 score of 0.63 with mBERT on the original English dataset, while the same model reported a 0.55 F1 score on the French translation. Testing on FrenchPolArg in the model transfer setting reported an F1 score of 0.48. We argue that the performance drop between the English and the French version of ElecDeb60to20 proves the impact that language shifts has on the component detection task. Moreover, we tried a data mixing approach by augmenting FrenchPolArg with data from ElecDeb60to20, delivering more accurate results (a macro F1 of 0.58 with mBERT). This demonstrates the tendence of language

models to learn textual patterns in the data rather than the task itself (Feger *et al.*, 2025). We conclude that a semi-automatic annotation workflow is possible and we believe that it would reduce the time and cost of the process. However, our results also confirm that the involvement of human annotators is necessary for a successful annotation campaign.

6 Conclusion and Future Works

This paper presents an annotation framework designed to produce a previously unavailable resource: a multilingual political dataset annotated to enable the study of the logical structures of arguments. The full annotation campaign is in progress, but a seed dataset composed of two languages (English and French) and partly annotated for argument components is already available. We plan on finalizing the annotation process in the next months, thus providing annotations for the full pipeline explained in this paper on all of our target languages, together with an evaluation that will follow Section 4. Rather than serving as an end goal, this dataset constitutes a crucial step in a thesis focused on the NLP-based analysis of argumentative reasoning within the political domain. We presented a state of the art focused on the reasons why this resource is necessary, namely the need for data annotated with the logical structure of the text. We also presented already existent resources, to highlight the lack of a logic-based multilingual dataset. We elaborated an annotation framework based on an abstract argumentation model, aiming at bridging the gap between formal models and LLMs. We integrated this workflow in an already started annotation campaign that involves expert annotators from SciencesPo Paris, and performed transfer learning studies on the subtask of detecting claims and premises, already described in a previously published paper.

Our next step will be implementing a logic-based extraction of arguments in political text. We will leverage the dataset we published as a ground truth against which evaluate large language models and test their reasoning skills based on the ASPIC+ argumentation framework. Parallel to this, the multilingual and multicultural aspect of the dataset enables the development of more robust and generalizable algorithms, as well as a comparative analysis across different languages and cultures.

Acknowledgements

I would like to thank my advisors Benoît Sagot and Chloé Clavel for their support in this project, as well as Emiliano Grossmann from the Center of European Studies at SciencesPo Paris for his help retrieving the data. This work was carried out at Inria Paris and I am grateful to the CLEPS infrastructure for providing resources and support.

References

ALLAWAY E. & MCKEOWN K. (2025). Evaluating defeasible reasoning in LLMs with DEFREASING. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 10540–10558, Albuquerque, New Mexico. DOI : 10.18653/v1/2025.naacl-long.529.

BAIN M., HUH J., HAN T. & ZISSERMAN A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, p. 4489–4493. DOI : 10.21437/Interspeech.2023-78.

EGER S., DAXENBERGER J., STAB C. & GUREVYCH I. (2018). Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In E. M. BENDER, L. DERCZYNSKI & P. ISABELLE, Éds., *Proceedings of the 27th International Conference on Computational Linguistics*, p. 831–844, Santa Fe, New Mexico, USA.

FANG X., LI Z., CHEN C. & LIAO B. (2025). LLM-ASPIC+: A Neuro-Symbolic Framework for Defeasible Reasoning. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2025)*, volume 413 de *Frontiers in Artificial Intelligence and Applications*, p. 1567–1574: IOS Press. DOI : 10.3233/FAIA250981.

FEGER M., BOLAND K. & DIETZE S. (2025). Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 23900–23915, Vienna, Austria. DOI : 10.18653/v1/2025.acl-long.1164.

GOFFREDO P., CHAVES M., VILLATA S. & CABRIO E. (2023). Argument-based detection and classification of fallacies in political debates. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 11101–11112, Singapore. DOI : 10.18653/v1/2023.emnlp-main.684.

GRAIFF C., CLAVEL C. & SAGOT B. (2026). Cross-lingual and cross-country approaches to argument component detection: a comparative study. In P. CHEN, V. ZOUHAR, H. HU, S. KHANUJA, W. ZHU, B. HADDOW, A. BIRCH, A. F. AJI, R. SENNRICH & S. HOOKER, Éds., *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, p. 149–161, Rabat, Morocco: Association for Computational Linguistics. DOI : 10.18653/v1/2026.mme-main.9.

HADDADAN S., CABRIO E. & VILLATA S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 4684–4690, Florence, Italy. DOI : 10.18653/v1/P19-1463.

HELWE C., CALAMAI T., PARIS P.-H., CLAVEL C. & SUCHANEK F. (2024). MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 4810–4845, Mexico City, Mexico. DOI : 10.18653/v1/2024.naacl-long.270.

KAZEMI M., YUAN Q., BHATIA D., KIM N., XU X., IMBRASAITE V. & RAMACHANDRAN D. (2023). Boardgameqa: A dataset for natural language reasoning with contradictory information.

KLIE J.-C., BUGERT M., BOULLOSA B., ECKART DE CASTILHO R. & GUREVYCH I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In D. ZHAO, Éd., *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, p. 5–9, Santa Fe, New Mexico: Association for Computational Linguistics.

LALWANI A., KIM T., CHOPRA L., HAHN C., JIN Z. & SACHAN M. (2025). Autoformalizing Natural Language to First-Order Logic: A Case Study in Logical Fallacy Detection. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, p. 132–147, Mumbai, India.

LAWRENCE J. & REED C. (2020). Argument mining: A survey. *Computational Linguistics*, **45**(4), 765–818. DOI : 10.1162/coli_a_00364.

LEI Y. & HUANG R. (2024). Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 13157–13173, Miami, Florida, USA. DOI : 10.18653/v1/2024.emnlp-main.730.

LI H., SCHLEGEL V., SUN Y., BATISTA-NAVARRO R. & NENADIC G. (2025). Large Language Models in Argument Mining: A Survey.

LUMER C. (2022). An Epistemological Appraisal of Walton's Argument Schemes. *Informal Logic*, **42**(1), 203–290. DOI : 10.22329/il.v42i1.7224.

MAYER T. *et al.* (2021). Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artificial Intelligence in Medicine*.

MODGIL S. & PRAKKEN H. (2014). The ASPIC + Framework for Structured Argumentation: a Tutorial. *Argument and Computation*, **5**(1), 31–62. DOI : 10.1080/19462166.2013.869766.

NUTE D. (1994). Defeasible logic. In D. GABBAY, C. HOGGER & J. ROBINSON, Éds., *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, p. 353–395. Oxford, UK: Oxford University Press.

RAMSHAW L. & MARCUS M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

REED C. & ROWE G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, **14**. DOI : 10.1142/S0218213004001922.

RENNARD V., SHANG G., GRARI D., HUNTER J. & VAZIRGIANNIS M. (2023). FREDSum: A dialogue summarization corpus for French political debates. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 4241–4253, Singapore. DOI : 10.18653/v1/2023.findings-emnlp.280.

RUIZ-DOLZ R., CHIU C.-J., CHEN C.-C., KANDO N. & CHEN H.-H. (2024). Learning strategies for robust argument mining: An analysis of variations in language and domain. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 10286–10292, Torino, Italia: ELRA and ICCL.

SCHAEFER R., KNAEBEL R. & STEDE M. (2022). On selecting training corpora for cross-domain claim detection. In G. LAPESA, J. SCHNEIDER, Y. JO & S. SAHA, Éds., *Proceedings of the 9th Workshop on Argument Mining*, p. 181–186, Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics.

STAB C. & GUREVYCH I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, **43**(3), 619–659. DOI : 10.1162/COLI_a_00295.

STAB C., MILLER T. & GUREVYCH I. (2018). UKP Sentential Argument Mining Corpus. In *Proceedings of the 9th Workshop on Argument Mining*.

TACHMAZIDIS I., BATSAKIS S. & ANTONIOU G. (2024). Benchmarking Defeasible Reasoning with Large Language Models – Initial Experiments and Future Directions.

TIEDEMANN J. & THOTTINGAL S. (2020). Opus-mt – building open translation services for the world. In A. MARTINS [ET AL.], Éd., *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 479–480, Switzerland: European Association for Machine Translation. Annual Conference of the European Association for Machine Translation , EAMT2020 ; Conference date: 03-11-2020 Through 05-11-2020.

TOLEDO-RONEN O., ORBACH M., BILU Y., SPECTOR A. & SLONIM N. (2020). Multilingual argument mining: Datasets and analysis. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 303–317, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.29.

WALTON D. & HANSEN H. (2013). Arguments from Fairness and Misplaced Priorities in Political Argumentation. *Journal of Politics and Law*, **6**(3), 78–94.

WALTON D., REED C. & MACAGNO F. (2012). *Argumentation Schemes*. Cambridge University Press. DOI : 10.1017/CBO9780511802034.

YEGINBERGEN A., ORONOZ M. & AGERRI R. (2024). Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 11687–11699, Bangkok, Thailand. DOI : 10.18653/v1/2024.acl-long.628.

A Table of Argument Schemes

Table 3: Argument schemes selected from (Walton *et al.*, 2012) and (Walton & Hansen, 2013), together with their explanation and formalization, partly taken from (Lumer, 2022).

Argument type	**Reasoning**
Argument from sign	- Major Premise: A sign S is observed. - Minor Premise: If sign S is observed, it is an indication of event E. - Conclusion: Event E is happening.
Argument from position to know	- Major Premise: Source S is in position to know fact F. - Minor Premise: S asserts that F. - Conclusion: F is true (or false).
Argument from expert opinion	- Major Premise: Source E is an expert in subject domain D containing proposition P. - Minor Premise: E asserts that P is true (false). - Conclusion: P is true (false).

Argument type	**Reasoning**
Argument from popular opinion	- Major Premise: Proposition P is widely believed by population E. - Minor Premise: E asserts that P is true (false). - Conclusion: P is true (false).
Argument from ignorance	- Major Premise: If A were true, A would be known to be true. - Minor Premise: A is not known to be true. - Conclusion: A is false.
Circumstantial ad hominem argument	- Argument Premise: A advocates argument α, which has proposition P as its conclusion. - Inconsistent Commitment Premise: A is personally committed to the opposite (negation) of P, as shown by commitments expressed in her/his personal actions or personal circumstances expressing such commitments. - Credibility Questioning Premise: A's credibility as a sincere person who believes in his own argument has been put into question (by the two premises above). - Conclusion: The plausibility of A's argument α is decreased or destroyed.
Argument from cause to effect	- Minor Premise (Cause Identification): Event A (the cause) occurs or exists. - Major Premise (Causal Link): Event A is known to produce Event B (the effect). - Conclusion: Therefore, Event B will occur.
Argument from correlation to cause	- Correlation premise: There is a positive correlation between A and B. - Conclusion: A causes B.
Argument from positive consequences	- Minor Premise: Value V is positive as judged by agent A. - Major Premise: If V is positive, it is a reason for A to commit to goal G. - Conclusion: V is a reason for A to commit to goal G.
Argument from negative consequences	- Minor Premise: Value V is negative as judged by agent A. - Major Premise: If V is negative, it is a reason for retracting commitment to goal G. - Conclusion: V is a reason for retracting commitment to goal G.
Slippery Slope Argument	- First step premise: A_0 is up for consideration as a proposal that seems initially like something that should be brought about. - Recursive premise: Bringing up A_0 would plausibly lead (in the given circumstances, as far as we know) to A_1, which would in turn plausibly lead to A_2, and so forth, through the sequence $A_2, \ldots$ A_n. - Bad outcome premise: A_n is a horrible (disastrous, bad) outcome. - Conclusion: A_0 should not be brought about.
Argument from analogy	- Premise 1: I have a goal G. - Premise 2: Doing G is similar to doing action X. - Premise 3: X has specific characteristics, such as being positive or negative. - Conclusion: Therefore, the same characteristics apply to action G.
Argument from (verbal) classification	- Major Premise: A has property F. - Minor Premise: F is classified as G. - Conclusion: A has property G.

Argument type	**Reasoning**
Argument from practical reasoning	- Major Premise: I have a goal G. - Minor Premise: Carrying out this action A is a means to realize G. - Conclusion: Therefore, I ought (practically speaking) to carry out this action A.
Argument from generalization	- Major Premise: Action A is usually performed in context C. - Minor Premise: Agent B performed action A. - Conclusion: Agent B is in context C.
Argument from example	- Major Premise: Example E leads to conclusion C. - Minor Premise: Example E is similar to situation S. - Conclusion: Situation S will have conclusion C.