# Cross-lingual and cross-country approaches to argument component detection: a comparative study

**Cecilia Graiff** [1], **Chloé Clavel** [1], **Benoît Sagot** [1]
[1] ALMAnaCH, Inria Paris
name.surname@inria.fr

## Abstract

Argument mining in multilingual settings has rarely been investigated, due to the lack of annotated resources and to the inherent difficulty of the task. We benchmark the performance of models on cross-lingual and cross-country argument component detection, focusing on political data from the US and France. To do so, we introduce FrenchPolArg, a corpus of argumentative political discourse in French, and we automatically translate already existing US-English resources. We benchmark three different cross-lingual and cross-country pipelines, and compare their results to find the best-performing one. We obtain promising results to be integrated in semi-automatic annotation workflows to reduce the time and cost of annotations.

## 1 Introduction

Two main gaps in argumentation mining research are the lack of annotated data and the poor generalizability of models across domains, languages, and datasets. Because of the high subjectivity of the task and the differences in argumentation styles, models tend to learn data rather than arguments (Feger et al., 2025). Moreover, the available datasets are mostly in English, and do not cover a broad range of topics. This paper aims at evaluating the performance of a cross-lingual pipeline for the task of argument component detection in a bilingual setting, thus contributing to the generalizability of argument retrieval by language models. We focus our analysis on political-domain data in the US and France. As can be imagined, this setting presents differences in both language and culture. While we are aware of the necessity of differentiating between the former and the latter, this paper will not investigate the relation between them, due to our choice of focusing on cross-lingual model generalizability rather than an accurate sociolinguistic analysis. To avoid an oversimplified use of the expression "cultural differences," we will refer to this aspect as "country-level." We further define our work with the following research questions (RQs):

**RQ 1: How much do language differences influence the model's performance?** We translate the English dataset ElecDeb60to20 and project the annotations, with the purpose of evaluating multilingual models on the translated version, thus assessing the performance drop. This approach allows us to focus on the cross-lingual aspect, and obtain a faithful estimation of how much adopting a cross-lingual pipeline impacts the model's performance. This experimental setting is crucial to assess the importance of one of the main problems of the current state of the art in argumentation mining research, namely the difficulty of learning and generalizing tasks between two languages.

**RQ 2: How much do country-level differences impact the performance of language models in the argument component detection task?** While the problem of robustness and generalizability is inherent to the argument mining task, the further nuance of country-level difference is rarely taken into account, even though some existing studies focus on cross-lingual or cross-domain problems (Eger et al., 2018; Yeginbergen et al., 2024a; Schaefer et al., 2022). However, we stress the relation between language, culture, and argumentation style, thus investigating its impact on model performance. We add a further layer to the first research question by increasing the difficulty of model generalizability, adding a change of culture to the change of language.

**RQ 3: How can we improve the cross-lingual and cross-country generalizability of language models for the argument component detection task and allow their reusability on different datasets?** One of the main challenges of argu-



*Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 149–161
March 28, 2026 

ment mining is the lack of available annotated data, especially in multilingual scenarios. Therefore, we deliver information about which ones among the chosen cross-lingual and cross-country pipelines help improve generalizability. This part of the work also leads the way to the creation of a semi-automatic annotation workflow to help the costly and difficult process of delivering new annotated datasets.

**Contributions** Our contributions are as follows:

1. We build FrenchPolArg, a corpus of French presidential debates and speeches, and annotate part of it to create a ground truth. We aim at providing a dataset for a domain for which, to the best of our knowledge, no annotated resource is currently available in French. Because non-English data is particulary scarce in the argument mining field, we believe that this resource addresses one of the most important gaps in the current research.

2. We automatically translate into French ElecDeb60to20 (Goffredo et al., 2023), a corpus of US presidential debates annotated for argument components. We project the original annotations of this corpus, thus obtaining an annotated resource of argumentative political text in French. We are aware that language is not the only difference between a French and an English corpus - cultural differences certainly play a role too. Our aim is to leverage these data for cross-lingual experiments.

3. We benchmark the performances of language models on cross-language and cross-country transfer, and compare the results to assess performance drop. Furthermore, we deliver information about the best performing pipeline to address language and country differences.

## 2 Related Work

Argument component detection is a subtask of argument mining that focuses on detecting claims and premises. Several studies have been published on the topic (Lawrence and Reed, 2020), but they mostly focus on a monolingual setting with no domain variation, leaving a research gap concerning multilingual and multicultural approaches. Because this task lies between the realms of argument mining and claim detection, we extend our literature review to related tasks such as cross-domain argument mining and claim detection.

**Cross-lingual transfer learning with masked language models** Cross-lingual transfer learning aims to enhance the performance of models on target languages by leveraging knowledge acquired from different source languages (Zhuang et al., 2021). Because some preliminary experiments with LLMs did not deliver good results, we chose to focus on BERT-based models. The better performance of BERT-based models on similar tasks, such as claim detection or fake news detection, is confirmed in the literature. Raza et al. (2024) benchmark BERT-based models and LLMs on the classification of fake news detection, and report F1 scores close to 90% for `roberta-base-uncased`, while the best performing LLM-based approach, which is fine-tuning Mistral-7B-v0.2, reports an F1 score of 80.23%. Azuma et al. (2025) perform an accurate comparative study of SVMs, BERT-like models, and LLMs (Mistral, LLaMA 3.2-3B, and Qwen3-4B) and determine that for specialized tasks, task-specific fine-tuning of a smaller, specialized model remains a more effective and computationally efficient approach than in-context learning or even parameter-efficient fine-tuning with larger, general-purpose generative models. Among the models leveraged for cross-lingual transfer learning, mBERT achieves very good results. mBERT follows the architecture of BERT, and was trained on with data from Wikipedia in 104 languages and without cross-lingual signal. Muller et al. (2021) perform a structural and behavioral analysis of the language-transfer capabilities of mBERT, concluding that this language model is composed by two sub-networks. The first one is a multilingual encoder, followed by a task-specific language-agnostic predictor. Wu and Dredze (2019) test mBERT on 5 different tasks (document classification, natural language inference, named entity recognition, part-of-speech tagging, and dependency parsing) from English to 38 target languages. Their results show that mBERT always achieves optimal results, sometimes state-of-the-arts results. Cross-lingual transfer for sequence labeling tasks is evaluated by García-Ferrero et al. (2022) and more specifically for argument mining by Yeginbergen et al. (2024a), as explained in the next paragraph.

**Cross-lingual argument mining** One of the first comprehensive experiments about multilingual argument component detection is (Eger et al., 2018). Similarly to the first part of our methodology, they (automatically and manually) translated into German, French, Spanish, and Chinese an English dataset of student essays and projected the annotations. While they obtained relatively good result, the inevitable performance drop is clearly visible. (Yeginbergen et al., 2024a) demonstrate that in the case of multilingual argument mining, data-transfer methods outperform model-transfer, where "model-transfer" denotes the use of a language model's experience in one language to generalize and apply it to another language. Their study is a further proof of the tendency of language models to learn data rather than the argument mining task itself. However, their results differ from prior findings on related sequence labeling tasks in the community (García-Ferrero et al., 2022). Our paper partially builds on their proposed approach by leveraging the automatic translation and annotation projection pipeline. While delivering different results from their paper, RQ3 partly shares their aim and structure. However, they work with the medical dataset AbsRCT (Mayer et al., 2020), consisting of randomized controlled trials retrieved from the MEDLINE database via PubMed search, whereas our dataset is political and presents a conversational argumentative context. Therefore, our works entails important differences that make our contribution unique, as no similar research exists on political and conversational data. It is important to mention that we introduce the cross-country differences, which we compare to cross-domain tasks, whereas Yeginbergen et al. (2024a) is focused on cross-lingual transfer. To do so, we build a new dataset of French political argument, leveraged as unseen test set. (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. The combination of different datasets is a strategy used in this paper as well, as it allows to augment the training size while decreasing the probability of the model overfitting on a single dataset. Their approach aims at finding the best dataset combination, hence their choice of experimenting only with RoBERTa. Differently from them, we include data mixing as an approach among others, and experiment with several models. Moreover, they work in a monolingual setting, while this paper merges the cross-lingual and the cross-country transfer, thus providing different insights.

**Cross-domain argument mining** While language models are able to perform cross-lingual transfer learning, it is worth noticing that a shift in language does not take into account country-related differences. For this reason, we consider the methods applied to cross-domain argumentation research to be pertinent to our work. (Daxenberger et al., 2017) were among the first to compare in-domain and cross-domain claim detection. More specifically, they compared user-generated web discourse such as blog posts, persuasive essays, online comments, Wikipedia talk pages, argumentative microtexts, and include the remaning categories as "various genres." Their results reveal a strong performance drop in cross-domain contexts. Moreover, they observe that feature-based approaches outperformed deep learning approaches, thus revealing that the presence of lexical indicators guides the model's choice. A similar result is achieved by (Alhamzeh et al., 2021), with slightly higher scores for traditional feature-based methods compared to DistilBERT. This paper also shows performance improvement when leveraging ensemble learning, and focuses on user-generated web content and student essays. However, while they obtain an F1 score of 0.79 on the cross-domain task, it is worth noticing that they train the model on a mix of both the corpora they investigate, thus not focusing on generalizability and allowing the model to still learn the dataset more than the task itself. In our work, we benchmark other approaches as well, to avoid biasing our interpretation of the cross-lingual transfer abilities of models. A more recent work on stance classification evidences the lack of cross-domain works, as well as the necessity for annotated cross-domain data (Yuan et al., 2024). While our setting is comparable to cross-domain under the technical point of view, the cross-country question that we are analyzing is underrepresented in the field.

## 3 Methodology

The cross-lingual aspect of this research is first focused on a bilingual study of English and French. We test three approaches on these two languages, namely (1) data transfer, (2) model transfer, and (3) data mixing. To the best of our knowledge, no annotated resource exists in French for the political domain. Therefore, we build FrenchPolArg, a French dataset of political discourse, and partially annotate it with argument components (claims

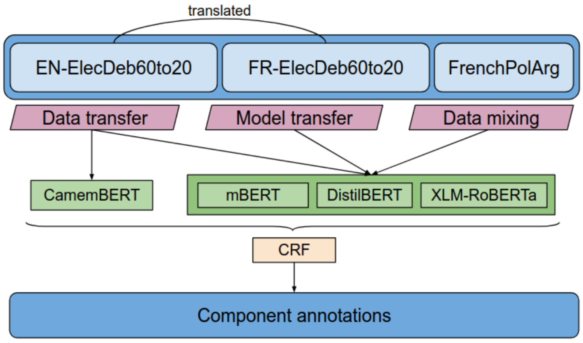


Figure 1: Pipeline of our work.

and premises) to build a ground truth. To perform cross-lingual experiments that can lead the way to a semiautomatic annotation workflow, we translate ElecDeb60to20 (Goffredo et al., 2023), a collection of presidential debates in the US from 1960 to 2020 annotated for argumentative components and relations. The implemented pipeline follows the approach described in (Yeginbergen et al., 2024b), based on machine translation (MT) and annotation projection. We translate the dataset into French, thereby generating parallel training sets used to fine-tune BERT-based classifiers for the task in the three settings mentioned above. We subsequently evaluate the model on the unseen real-world data contained in FrenchPolArg, to assess model performance and to perform a comparative analysis of the approaches. The data and the code are available here. An overview of the pipeline used for this paper is available in Figure 1. In the next paragraphs, we will present the used approaches and the dataset collection, and later we will explain the experimental setting and fine-tuning steps.

### 3.1 Followed approaches

We adopt three different approaches, based on:

1. **Data transfer**: Data transfer consists in generating training data by translating ElecDeb60to20 (Goffredo et al., 2023) into French and projecting the annotations, thus obtaining FR-ElecDeb60to20. We test this setting with both multilingual and monolingual models. While we also leverage multilingual models in the model transfer strategy, we consider the latter fundamentally different, because the models are trained and tested on data in the same language.

2. **Model transfer**: Differently from the data transfer approach, model transfer consists in leveraging the cross-lingual transfer abilities of the model itself, rather than applying a pipeline of translation and projection with the aim of generating new data containing the same knowledge in another language. We leverage multilingual pre-trained language models. We use the standard English version of ElecDeb60to20 as a training set, and test the model on its French translation to investigate the impact of language differences. We then test on FrenchPolArg to add the further layer of country differences.

3. **Data mixing**: We progressively augment the French and English versions of ElecDeb60to20 with data from FrenchPolArg, in order to iteratively improve the results. We consider the limited size of the annotated portion of FrenchPolArg as a limitation. Therefore, the data augmentation approach allows us to have sufficient data for training a BERT-based model, while diminishing the probability of the model overfitting on FrenchPolArg.

### 3.2 Dataset creation

#### 3.2.1 FrenchPolArg

**Dataset collection** FrenchPolArg is composed by the transcripts of 8 presidential debates broadcasted on television in 1974, 1981, 1988, 1995, 2007, 2012, 2017, and 2022. The debates are publicly available online. When a transcript already existed, it was scraped; alternatively, the YouTube video of the debate was transcribed and diarised with the WhisperX model (Bain et al., 2023), which provides fast automatic speech recognition (70x real time with large-v2) and was chosen over Whisper because it includes speaker diarisation. Moreover, transcripts of French presidential speeches and declarations were scraped from the official website of the Élysée and of Vie Publique. Because this is not a conversational setting, we choose not to focus on it for this paper. As the speeches are uploaded on the website in PDF form with plenty of noisy text, the scraped content was thoroughly preprocessed, both automatically and manually. As far as we know, the only similar resource is FREDSum (Rennard et al., 2023), a corpus of French political debate annotated for summarization. While the data partially overlap, our contribution differs from FREDSum because it covers both debate and speeches, it does not entail a separation based on topics but only on debates, and it is partially annotated with argument components, as will be explained in the next section. Differently from FREDSum, we chose to not split the dataset based on the debate topics, as we aim at having a dataset that spans more broadly in order to better evaluate generalizability in this paper.

**Annotation of the test set** Due to the length and complexity of the annotation task, it was not possible to annotate the whole FrenchPolArg. This resource problem led us to choose to annotate a part of the dataset to use as ground truth to evaluate model generalizability. To minimize diachronic bias, we annotate the first, last, and middle available debate, namely the ones between François Mitterrand and Valéry Giscard d'Estaing (1981), Jacques Chirac and Lionel Jospin (1995), and Emmanuel Macron and Marine Le Pen (2022). The annotations were performed by a 25-year-old French-speaking female annotator, who also is one of the authors of this paper.

We define a set of arguments $A$, composed of a set of premises $P$ and a set of claims $C$, where the premises have the function of supporting or attacking the claims. To take into account the faulty logical reasoning often present in natural language arguments, we allow the existence of claims that are not supported by premises, but not of premises that do not support any claim. We choose to not limit the amount of accepted claims to only one, to allow more complex argument structures. The annotation of argumentative components was performed with the INCEpTION platform and focused on the token-level. We annotated 27,299 tokens (1509 sentences) following the Beginning-Inside-Outside (BIO) scheme (Ramshaw and Marcus, 1995), which marks tokens as beginning, inside, or outside an argument component.

### 3.2.2 Translated ElecDeb60to20

**Description of the original dataset** To the best of our knowledge, the only political dataset annotated for argument components and relations is ElecDeb60to20 (Goffredo et al., 2023). This dataset contains the debates between presidency and vice-presidency candidates of the United States of America from 1960 until 2020. The raw data was originally scraped from the Commission on Presidential Debates (CPD). A first round of annotations was performed on the data from 1960 to 2016; later, the dataset was extended to comprehend the debates until 2020. The annotation task was performed manually, reaching a high inter-annotators agreement with values such as a Krippendorf's alpha of 0.757 for the second annotation round. Apart from its political nature, we chose this dataset because the annotation scheme is compatible with ours, explained in the next paragraphs. Moreover, the good results obtained by ElecDeb60to20 in previous research works in the argument mining field deemed it a suitable choice. The use of ElecDeb60to20 required accurate preprocessing. The part of the dataset annotated for argumentative components presented several duplicates, that we removed automatically. Duplicates with different annotations were removed manually and not automatically, in order to avoid the potential loss of information and to ensure the quality of the final dataset.

**Translation and annotation projection** Following the pipeline implemented by (Yeginbergen et al., 2024a), we automatically translate ElecDeb60to20 into French [1]. We leverage Opus-MT (Tiedemann and Thottingal, 2020), a set of language models pre-trained on the machine translation task. Even though Yeginbergen et al. (2024a) observe a slightly better performance of DeepL compared to Opus-MT, we choose the latter model because it is open source. While our focus in this paper is bilingual, we aim at extending this work to a multilingual setting, thus motivating the necessity of a translation service that covers a wide range of languages. Therefore, we make the translated corpus available in French. The original annotations of components were projected with SimAlign (Sabet et al., 2021), particularly suited to the task of aligning sequences in different languages, where the structure of sentences does not necessarily match. Here as well, the choice is motivated by the good results reported by Yeginbergen et al. (2024a) and García-Ferrero et al. (2022) on this task. For further clarity, we present an overview of the used datasets in Table 1.

[1] We also provide translations into German, Spanish, and Italian. This paper focuses on a bilingual study, but we plan to expland it to a multilingual setting.

| Dataset Name | Dataset Description | Size (tokens) |
|---|---|---|
| EN-ElecDeb60to20 | Annotated US presidential debates in English. (Goffredo et al., 2023). | 707,976 |
| FR-ElecDeb60to20 | French translation of EN-ElecDeb60to20. | 741,835 |
| FrenchPolArg | Annotated French presidential debates. | 68,392 |

Table 1: Overview on the used datasets.

### 3.3 Model training

For each of the above presented settings, we test several different configurations. In this section, we will present the tasks and the data splitting procedure; for an overview of the selected architectures, please refer to Appendix B.

**Fine-tuning task** The models are fine-tuned on a multiclass token classification task, following the adopted BIO annotation scheme to represent components' boundaries. Given a sentence $S$ consisting of $n$ tokens, our goal is to predict for each token $t_i$ if it is a claim ($y_i = 1$), a premise ($y = 2$), or not argumentative ($y_i = 0$). Therefore, our task is originally formulated as a 3-classes token classification task, with the aim of labeling tokens as claim, premise, or non argumentative. However, because we adopt the BIO annotation scheme, our classes are expanded to include the begin (B), inside (I), and outside (O) tags. Moreover, we want to exclude the tags generated by the tokenizer, such as [SEP], [CLS], [PAD]. Therefore, our original 3-classes problem is expanded to a 6-classes problem: B-Claim, I-Claim, B-Premise, I-Premise, O for non argumentative tokens, and X for the tokenizers' tags. Because the model's performance on classifying the tags reached top accuracy (more than 0.99), we consider it ininfluential and evaluate our model on the 5 remaining classes.

**Data splitting** We split the monolingual datasets respecting class proportions, as explained in detail in appendix A. For the multilingual dataset, we split equally the train, dev, and test set between the two languages to avoid cross-lingual duplicates. Because both English and French are not considered low-resource languages, we assume that we can rely on the transfer learning capabilities of multilingual models. However, we are aware of the bias caused by the fact that the amount of English in their training data is superior to any other language. In the data mixing setting, the training dataset is built by adding samples from the target dataset to FR-ElecDeb60to20. We avoid increasing the training dataset with more than 20% of FrenchPolArg to not decrease too much the test set size. We call the thus obtained subsets FrenchPolArg-S1 and FrenchPolArg-S2, where the first one is used to increase the train dataset size, and the second one is used as test dataset. We call the increased training datasets EN-aug, which consists of EN-ElecDeb60to20 augmented with FrenchPolArg-S1, and FR-aug, which consist of FR-ElecDeb60to20 augmented with FrenchPolArg-S1. All experiments are tested on FrenchPolArg-S2.

### 3.4 Evaluation

#### 3.4.1 Evaluation of the translation and projection pipeline

We evaluate the translation and projection pipeline by manually correcting a randomly selected sample of 100 sentences (1556 tokens) and checking the token distribution. We report in Table 2 the difference in the token distribution between the original and backtranslated English versions of ElecDeb60to20. The backtranslated version is obtained by automatically translating ElecDeb60to20 into French and projecting the annotations, and then repeating the same process from French to English. These absolute and percentual values are used as an evaluation of the accuracy of the projections, together with the manual evaluation of a random sample of 100 sentences (1556 tokens). We focus on the B-tokens, which represent the amount of claims and premises in both datasets. B-Claim and B-Premise present a decrease, meaning that the annotation missed some tags, but the percentual difference in the whole dataset is very small (around 0.25%). Parallel to this quantitative analysis, we perform a qualitative one by analizing 100 random sentences (1556 tokens). We report only two minor mistakes in the projection, and only three cases where a too literal translation hinders the understanding of FR-ElecDeb60to20 (such as "donner un coup de pied" for "kicking in"). We are aware that no perfect model exists, and consider the obtained results to be satisfactory enough to ensure the overall correctness of the translation projection pipeline. Moreover, as our aim is to test this configuration, we do not correct the errors manually, in order to obtain an unbiased evaluation of our approach.

| Tag | Original | Backtranslated | Difference (%) |
|---|---|---|---|
| O | 226,333 | 241,982 | +4.2% |
| B-Claim | 12,147 | 10,458 | -0.26% |
| I-Claim | 139,457 | 134,204 | -0.25% |
| B-Premise | 10,743 | 9,100 | -0.26% |
| I-Premise | 145,506 | 123,473 | -3.48% |

Table 2: Differences in the token distribution between the original and backtranslated English version of ElecDeb60to20. The percentage indicates the change in relation to the whole corpus. We use these numbers to evaluate the quality of annotation projection.

**Evaluation of the cross-lingual transfer experiments** The models are evaluated with micro and macro F1, which are common metrics chosen for sequence classification tasks. While macro F1 weighs each class equally, micro F1 weighs each sample equally, thus allowing us to perform a proper evaluation even though the classes are not equally distributed. Because the differences between micro and macro F1 were minimal, we report only the macro scores in the paper. Full tables comprehending the micro F1 scores are available in Appendix D. To avoid that the model's ability of recognizing non-argumentative tokens (labeled as O) biases the metric of model accuracy, we also report the specific F1 scores for claims and premises in Appendix D.

## 4 Results

With our experiments, we aim at answering the following research questions:

**RQ 1: How much do language differences influence the model's performance?** Before testing the three previously described approaches, we extensively experiment with EN-ElecDeb60to20 and FR-ElecDeb60to20 to investigate the impact of the language shift while eliminating country-related confounders. We present the results in Appendix D. As expected, we note a performance drop: the task reports an F1 score of 0.637 (rounded to 0.63) on the English dataset with the best performing model (mBERT), and of 0.46 on the French dataset, suggesting that the change in language has an important impact on the model's performance. Interestingly, we do not report significant differences among the models' performances, including CamemBERT, the only non-multilingual one. While the absolute values are lower, the proportional decrease described in this paper does not show important differences.

We also run the same experiment on the backtranslated version of the English corpus. The results are reported in Appendix E. We notice that while the translation to French involved a performance decrease, testing mBERT on the backtranslation does not present significant differences in performance compared to the French version. Hence, we argue that the translation, while not perfect, is not enough to explain all of the performance drop. Because the evaluation of the translation and projection pipeline was satisfactory, we exclude this possible cause, and hypothesize that the difficulty of cross-lingual transfer of the argument mining task be the reason. Therefore, we are convinced that the model's robustness needs to be improved also under the cross-lingual point of view.

To further prove this point, we test a model trained on EN-ElecDeb60to20 on its French translation, which causes a decrease in performance from 0.637 (rounded to 0.63) to 0.559 (rounded to 0.55). The results are reported in Table 9. These results confirm the scarce ability of language models to generalize to different languages in the argument mining task, and deliver further proof of the impact of language shifts on model performance, thus answering our RQ 1.

| Language | Model | Macro F1 |
|---|---|---|
| EN | mBERT | 0.63 |
| EN | XLM-RoBERTa | 0.58 |
| FR | mBERT | 0.46 |
| FR | XLM-RoBERTa | 0.46 |
| FR | CamemBERT | 0.45 |
| FR | DistilBERT | 0.44 |

Table 3: Experiments on the original and the translated version of ElecDeb60to20. All experiments present the same language in the training and test set. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

| Model | Macro F1 | F1 Premise | F1 Claim |
|---|---|---|---|
| mBERT | 0.55 | 0.51 | 0.46 |
| XLM-RoBERTa | 0.39 | 0.15 | 0.29 |

Table 4: Results of the models trained on EN-ElecDeb60to20 and tested on FR-ElecDeb60to20. We use the base, multilingual, uncased version of mBERT and XLM-RoBERTa.

**RQ 2: How much do country-level differences impact the performance of language models in**

**the argument component detection task?** To include the country-level differences, we test the pipeline on FrenchPolArg, which is in the same language as FR-ElecDeb60to20, but presents a different culture. Our experiments can be seen in Table 5, while a more extensive report containing both micro and macro F1 score can be found in appendix D. We notice that the performance drop is only slightly superior to the experiments presented in Table 3, with mBERT scoring 0.55 macro F1 when tested on FR-ElecDeb60to20, and 0.50 on FrenchPolArg. While we still are in the political domain, we hypothesize that the argumentation styles might differ from one country to the other. Therefore, we interpret our experiments as a proof that while country-level differences can cause a bias, the main issue to address in the field concerns the cross-lingual differences.

**RQ 3: How can we improve the cross-lingual and cross-country generalizability of language models for the argument component detection task and allow their reusability on different datasets?** To answer this research question, we compare the three different approaches described in this paper (model transfer, data transfer, and data mixing) and test them in the above described configurations. We present a summary of all the conducted experiments in Tables 5 and 6, while a more extensive report containing both micro and macro F1 score can be found in appendix D. Differently from the results of Yeginbergen et al. (2024a), we notice that model transfer delivers better results than data transfer, with 0.50 as best results versus 0.43.

Among the model transfer experiments, mBERT scores significantly better than the other models, while still reporting a performance drop compared to the test on EN-ElecDeb60to20 reported in Table 3. We also note the very poor performance of XLM-RoBERTa, which is the only model to score better in the data transfer approach, but delivers worse results than any other model in both approaches. This result is surprising, because XLM-RoBERTa is usually reported to score better results on multilingual tasks; however, we hypothesize that the small size of the dataset might have had an impact on the performance of XLM-RoBERTa, while smaller models manage to deliver better results. We present further details about the chosen models in Appendix B, but because of time- and resource-related limitations, we could not investigate this aspect any further.

| Model | Macro F1 |
|---|---|
| ***Model transfer*** | |
| mBERT | 0.50 |
| distilBERT | 0.49 |
| XLM-RoBERTa | 0.35 |
| ***Data Transfer*** | |
| CamemBERT | 0.40 |
| mBERT | 0.42 |
| distilBERT | 0.43 |
| XLM-RoBERTa | 0.40 |
| ***Data+Model Transfer*** | |
| mBERT | 0.47 |

Table 5: All experiments and models. The models are tested on FrenchPolArg and trained respectively on EN-ElecDeb60to20 for the model transfer setting, FR-ElecDeb60to20 for the data transfer setting, and the multilingual dataset FR+EN-ElecDeb60to20 for the data+model transfer. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT.

Our results show that the data transfer approach delivers relatively poor results for all models. The best performing model in this scenario is distilBERT (F1 score of 0.43), but the difference with BERT (F1 score of 0.42) is negligible. We notice that there are no significant differences between the performances of monolingual models such as CamemBERT, and multilingual models.

Based on these results, we hypothesize that argumentative discourse is such a language- and culture-dependent process that even a correct translation significantly decreases accuracy. While the proposed approaches are not perfect, we suggest that data mixing as the best performing one with a 0.585 macro F1 score could be applied to unseen data in order to provide a first batch of annotations, to later revise manually. This semi-automatic workflow would significantly reduce the annotation cost and time.

## 5 Conclusion and Future Work

In this paper, we present an original French annotated resource. In addition, we provide the translated version of ElecDeb60to20 into French. We then test cross-lingual and cross-country generalizability, and provide information on the best-working approach. Therefore, we provide new resources in a severely lacking domain, and deliver information about the open problem of generalizability. For the future direction of this work, we

| Train dataset | Model | Macro F1 score |
|---|---|---|
| ***Data Mixing*** | | |
| EN-aug | XLM-Roberta | 0.51 |
| FR-aug | XLM-Roberta | 0.44 |
| EN-aug | mBERT | 0.58 |
| FR-aug | mBERT | 0.47 |
| EN-aug | distilBERT | 0.55 |
| FR-aug | distilBERT | 0.43 |
| FR-aug | CamemBERT | 0.43 |

Table 6: All experiments and models in the data mixing setting. We use the base, multilingual, cased version of DistilBERT, mBERT, and XLM-RoBERTa. EN-aug refers to EN-ElecDeb60to20 augmented with 20% of the data from FrenchPolArg, and FR-aug refers to the same augmentation applied to FR-ElecDeb60to20. All models have the addition of a CRF as last layer and all models are tested on the remaining 80% of FrenchPolArg.

believe that increasing the provided resources and their quality is essential. We do not exclude to implement further strategies to diminish the impact of the small dataset size, such as k-fold cross validation. Moreover, it would be very important to be able to further investigate the lower performance of XLM-RoBERTa on this dataset, which was not possible for this paper because of time and resources. Most importantly, we planned an annotation campaign to expand our dataset and widen the experimental setting. We then plan to extend the experiments of this paper to include more data, thus increasing the reliability of the results, even though we work on strategies for scarce data settings. We are interested in merging the presented pipelines with manual annotation workflows, thus reducing time and the cost of annotations. A further planned approach is to extend the translations to include other languages, such as German, Italian, and Spanish.

## 6 Limitations

While this work offers an insight into cross-lingual and cross-country generalizability, it entails several limitations. Argument mining established a clear definition of arguments as units of speech composed by claim and premises, connected by support or attack relations; however, the annotation task itself is more complicated, because real-world data struggle to match the mathematical formalizations of arguments. Moreover, FrenchPolArg was annotated by one single person, due to the length, complexity, and cost of the process. However, this paper aims at delivering a baseline for further research in scarce data settings. This explains another limitation of this paper, namely the dataset size. To avoid that this limitation would bias our results, we relied on an already existent bigger dataset, namely ElecDeb60to20. Despite this measure, the tests performed on FrenchPolArg rely on a smaller test dataset size, which possibly biases the results, a limitation to be addressed in the future works described above.

## References


Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2021. A stacking approach for cross-domain argument identification. In *Database and Expert Systems Applications*, pages 361–373, Cham. Springer International Publishing.

Daichi Azuma, René Meléndez, Michal Ptaszynski, Fumito Masui, Lara Aslan, and Juuso Eronen. 2025. Svm, bert, or llm? a comparative study on multilingual instructed deception detection. *AI*, 6(9).

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

CPD. The commission on presidential debates. `https://debates.org/`. Accessed: 2025-10-03.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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

Marc Feger, Katarina Boland, and Stefan Dietze. 2025. Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2024. Fake news detection: Comparative evaluation of bert-like models and large language models with generative ai-annotated data. *Preprint*, arXiv:2412.14276.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. FREDSum: A dialogue summarization corpus for French political debates. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253, Singapore. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *Preprint*, arXiv:2004.08728.

Robin Schaefer, René Knaebel, and Manfred Stede. 2022. On selecting training corpora for cross-domain claim detection. In *Proceedings of the 9th Workshop on Argument Mining*, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 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.

Vittorio Torri and Francesca Ieva. 2023. Polimi at clinkart: a conditional random field vs a bert-based approach.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024a. Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11687–11699, Bangkok, Thailand. Association for Computational Linguistics.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024b. Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques. *Preprint*, arXiv:2407.03748.

Jiaqing Yuan, Ruijie Xi, and Munindar P. Singh. 2024. A benchmark for cross-domain argumentative stance classification on social media. *Preprint*, arXiv:2410.08900.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

## A Data splitting

We divide ElecDeb60to20 into a training and test set of respectively 90% and 10% of the dataset. We further split the training set into a train (90%) and evaluation set (10%). We thus obtain three subsets, namely training (76,5%), validation (13.5%), and test set (10%). We opt for this splitting format to try to keep proportions as accurate as possible when testing the experiments on ElecDeb60to20 or on FrenchPolArg. To avoid unbalanced data splits, we use the `stratify` parameter of `train_test_split`. We thus divide our data equally between the 5 classes, obtaining the statistics shown in Table 7.

| | train | dev | test |
|---|---|---|---|
| B-Claim | 18,829 | 3,415 | 2,446 |
| I-Claim | 218,449 | 39,277 | 28,731 |
| B-Premise | 16,851 | 2,955 | 2,176 |
| I-Premise | 231,777 | 40,391 | 30,394 |
| O | 389,718 | 68,069 | 50,768 |

Table 7: Dataset statistics.

For the data transfer setting, we split ElecDeb60to20 into a train and a validation set of respectively 90% and 10%, respecting class proportions. We then leverage FrenchPolArg as a test set.

## B Model selection and architecture

After prior unsuccessful experiments with the default architecture, we fine-tune BERT-based models with a Conditional Random Field (CRF) as last layer, following (Goffredo et al., 2022). CRFs are often used in tasks that rely on BIO tags, because they ensure that predicted sequences respect the structural constraints of the annotation scheme. We give further details about the probability computation in appendix C. Following this architecture, we train the following models:

**mBERT**: We fine-tune `bert-base-multilingual-cased` for token classification.

**MultilingualDistilBERT**: We fine-tune `distilbert-base-multilingual-cased`, a distilled version of BERT. This model is often used on smaller datasets, because it retains performance while diminishing computational costs.

**XLM-RoBERTa:** This model was trained on 2.5 TB of newly created clean CommonCrawl data in 100 languages and reported very good results on cross-lingual tasks, even for low-resource languages (Conneau et al., 2020). As we face the issue of data scarcity, we hypothesize that a model performing well in a domain characterized by the presence of very little data could work well in our context.

**CamemBERT:** We test `camembert-base` (Martin et al., 2020) on the French portion of the dataset. We aim at exploring the data-transfer functionality and hypothesize that training a French model on translated data could lead to better results. We cross-check the reliability of our approach by testing CamemBERT on the English portion of the dataset, expecting a performance drop. However, we are aware that, despite being a French language model, CamemBERT has seen English data in its training phase, and therefore partially knows how to handle English words.

## C Use of CRFs

Differently from a standard linear classification layer, CRFs model the conditional probability of the classes as follows:

$$P(y \mid x) \propto \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)\right)$$

where $x$ is the vector of tokens observations) that form the sequence, $y$ is the vector of labels (states) over the tokens, $i$ is an index over the sequence tokens, $n$ is the length of the sequence, $j$ indexes the feature functions $f_j$ and $\lambda_j$ are the parameters to be learnt. (Torri and Ieva, 2023)

## D Complete evaluation metrics

We present in this section the complete evaluation metrics for the performed experiments. Specifically, Tables 8, 9, 10 and 11 present the micro and macro F1 scores for the overall performance of the model, whereas Table 12 reports the single F1 scores for Claim and Premise in the experiments performed on EN-ElecDeb60to20 and FR-ElecDeb60to20.

| Language | Model | Macro F1 | Micro F1 |
|---|---|---|---|
| EN | mBERT | 0.63 | 0.61 |
| EN | XLM-RoBERTa | 0.58 | 0.59 |
| FR | mBERT | 0.46 | 0.44 |
| FR | XLM-RoBERTa | 0.46 | 0.50 |
| FR | CamemBERT | 0.45 | 0.46 |
| FR | DistilBERT | 0.44 | 0.45 |

Table 8: Experiments on the original and the translated version of ElecDeb60to20. All experiments present the same language in the training and test set. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

| Model | Macro F1 | Micro F1 | F1 Premise | F1 Claim |
|---|---|---|---|---|
| mBERT | 0.55 | 0.56 | 0.51 | 0.46 |
| XLM-RoBERTa | 0.39 | 0.42 | 0.15 | 0.29 |

Table 9: Results of the models trained on EN-ElecDeb60to20 and tested on FR-ElecDeb60to20. We use the base, multilingual, uncased version of mBERT and XLM-RoBERTa.

| Model | Macro F1 | Micro F1 |
|---|---|---|
| ***Model transfer*** | | |
| mBERT | 0.50 | 0.53 |
| XLM-RoBERTa | 0.41 | 0.43 |
| ***Data Transfer*** | | |
| CamemBERT | 0.44 | 0.46 |
| mBERT | 0.42 | 0.41 |
| XLM-RoBERTa | 0.38 | 0.37 |
| ***Data+Model Transfer*** | | |
| mBERT | 0.47 | 0.47 |

Table 10: All experiments and models. The models are tested on FrenchPolArg and trained respectively on EN-ElecDeb60to20 for the model transfer setting, FR-ElecDeb60to20 for the data transfer setting, and the multilingual dataset FR+EN-ElecDeb60to20 for the data+model transfer. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT.

| Train dataset | Model | Macro F1 | Micro F1 |
|---|---|---|---|
| ***Data Mixing*** | | | |
| EN-aug | XLM-Roberta | 0.52 | 0.54 |
| FR-aug | XLM-Roberta | 0.44 | 0.52 |
| EN-aug | mBERT | 0.58 | 0.59 |
| FR-aug | mBERT | 0.45 | 0.51 |
| EN-aug | distilBERT | 0.57 | 0.59 |
| FR-aug | distilBERT | 0.43 | 0.50 |
| FR-aug | CamemBERT | 0.44 | 0.50 |

Table 11: All experiments and models in the data mixing setting. We use the base, multilingual, cased version of DistilBERT, mBERT, and XLM-RoBERTa. EN-aug refers to EN-ElecDeb60to20 augmented with 20% of the data from FrenchPolArg, and FR-aug refers to the same augmentation applied to FR-ElecDeb60to20. All models have the addition of a CRF as last layer and all models are tested on the remaining 80% of FrenchPolArg.

| Language | Model | F1 Claim | F1 Premise |
|---|---|---|---|
| EN | mBERT | 0.588 | 0.536 |
| EN | XLM-RoBERTa | 0.487 | 0.420 |
| FR | mBERT | 0.369 | 0.333 |
| FR | XLM-RoBERTa | 0.338 | 0.290 |
| FR | CamemBERT | 0.341 | 0.315 |
| FR | DistilBERT | 0.338 | 0.313 |
| FR+EN | mBERT | 0.501 | 0.449 |

Table 12: Experiments on the original and the translated version of ElecDeb60to20. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

## E Backtranslation

Table 13 presents the results of the argument component classification experiments performed with mBERT on the backtranslated English version of ElecDeb60to20. We chose mBERT as it was the best performing dataset in the previously tested settings.

| Macro F1 | Micro F1 | F1 Premise | F1 Claim |
|---|---|---|---|
| 0.47 | 0.46 | 0.37 | 0.34 |

Table 13: Results of the models trained on the back-translation into English of ElecDeb60to20. We use the base, multilingual, uncased version of mBERT, with the addition of a CRF as last layer.

## F Technical setup

We test our models in several configurations, with 5 epochs delivering the best results. We implement the AdamW optimizer. All models were trained on the cluster of our institution, using H100 GPUs.