Relex: A Common Denominator for Connectives and Discourse Relations in French

Fatou Sow[1,2], Yannick Toussaint[1], Mathieu Constant[2]

[1]LORIA, Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
[2]ATILF, Université de Lorraine, CNRS, F-54000 Nancy, France
{fatou.sow, yannick.toussaint}@loria.fr, mathieu.constant@univ-lorraine.fr

Abstract

The lack of sufficiently large French resources linking discourse connectives to the relations they express in context hinders the training of models that map connectives to their discourse relations. We address this gap by introducing two complementary datasets: a large-scale, semi-automatically annotated corpus and a manually validated dataset. Both resources annotate connectives and their discourse relations according to the French lexicon LEXCONN. Relying on the large-scale corpus, we train **Relex**, a CamemBERT-based model fine-tuned to predict, among 19 relation types, the relation expressed by a connective. Despite being trained on a fixed inventory of connectives, Relex extends to previously unseen connectives and achieves an F1 score of 0.59. The datasets and the trained model are publicly available [1].

1 Introduction

Connectives contribute to discourse coherence as they express relations between discourse units. Because they structure discourse, connectives and relations have been studied in linguistics and formalized through frameworks such as RST (Mann et al., 1989) or SDRT (Lascarides and Asher, 2007) while in natural language processing (NLP), they are leveraged to solve natural language inference (Chan et al., 2023) or question-answering tasks (Miao et al., 2024).

Several multilingual projects focus on the creation of annotated resources or the training of neural networks to extract these relations. In French, lexicons (Roze et al., 2012) and corpora (Péry-Woodley et al., 2011; Danlos et al., 2012) have been developed, but gathering sufficient data for training models remains a challenge for discourse-related tasks.

In this context, we present a semi-automatically annotated dataset of discourse connectives and relations. We use this dataset to train Relex, a CamemBERT model (Martin et al., 2020) fine-tuned to map a connective to the relation it expresses. By processing these connectives across diverse contexts, our model generalizes discourse relations from the data rather than simply memorizing a static mapping. Consequently, it can perform classification on connectives absent from the training data.

Our main contributions are as follows:

- An automatically annotated training dataset of French discourse connectives and their relations in context.
- Relex, a fine-tuned CamemBERT model that generalizes discourse relations to perform relation classification on connectives unseen during training.
- A manually annotated evaluation dataset of connectives absent from the training data.

2 Related Work

Connectives and the relations they express are generally represented in corpora or lexicons. For instance, the Penn Discourse Treebank (Prasad et al., 2018) (PDTB) is an English corpus annotating the arguments of a connective and the relation between those arguments. This framework has then been adapted for use in various languages such as Turkish (Zeyrek and Kurfalı, 2017) or Czech (Synková et al., 2024). Alongside corpora, lexicons such as DiMLex (Stede and Umbach, 1998) and LEXCONN (Roze et al., 2012) have been constructed based on specific theoretical criteria to manually identify connectives and their relations. In particular, LEXCONN chooses the relations defined by the SDRT framework. While these lexicons often include textual examples, they ultimately provide static, out-of-context representations of connectives and their associated relations.

The creation of these resources enables the train-

[1]https://github.com/FatouSow/Relex

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 636–648
August 2–5, 2026.

ing of predictive models, driving initiatives like the CoNLL-2015 shared task (Xue et al., 2015), which focuses on discourse parsing tasks such as connective identification and relation classification between discourse units. While CoNLL-2015 resources are limited to English data, the DISRPT shared tasks (Braud et al., 2025) are based on multilingual discursive datasets. However, the resulting models separate connective and relation identification and notably within the DISRPT framework, the connective identification task remains unavailable for French.

Additionally, these tasks can be merged to solve implicit relation classification between discourse units. The goal is to train the model to predict connectives in order to improve the prediction of implicit relations that do not rely on obvious lexical hints (Wu et al., 2023). These systems, however, do not focus on predicting the specific relation expressed by a connective in context

In French, other than LEXCONN, Laali and Kosseim (2017b) present a lexicon mapping LEXCONN connectives to PDTB relations. In addition, the current version of the French Discourse Treebank (FDTB) (Danlos et al., 2015) includes the manual identification of connectives while discourse units and SDRT relations are annotated in the ANNODIS corpus (Péry-Woodley et al., 2011). To the best of our knowledge, there is no existing French corpus annotating both connectives and discourse relations, nor a model designed to automatically annotate the relation of a connective in context.

3 Data

The goal of our work is to predict the relation expressed by a connective in a textual sequence, as illustrated in Example (1). The connective is isolated between special tags to draw the attention of the trained system to the target connective. Consequently, the model learns to map the connective to a relation in diverse contexts and ultimately generalizes the underlying relation itself. Therefore, a large amount of annotated data for connectives and their relations is essential to fine-tune a BERT model.

(1) ***[MARKER] Peu avant de [/MARKER]*** *mourir, Mio a promis à son mari qu'elle reviendrait à la saison des pluies.* → **Narration**
(*English*: ***Shortly before*** *she died, Mio promised her husband that she would return during the rainy season.*)

The training data was collected in two steps. First, we evaluated a corpus semi-automatically annotated for connectives, as explained in Section 3.1. Second, in Section 3.2, we automatically added annotations of the discourse relations expressed by the connectives. Finally, in order to evaluate our model's generalization on new data, we implemented a protocol to build a manually annotated evaluation dataset, as detailed in Section 3.3.

3.1 Validation of annotated connectives

Data	Token	Sentence
Le Monde	32,639,407	1,614,898
Le Monde Diplomatique	2,027,795	76,033
Wikiconflit	501,366	23,343

Table 1: Distribution of tokens and sentences of the raw corpus

We gained access to automatically collected French data from newspapers *Le Monde* and *Le Monde Diplomatique* as well as comments from the platform *Wikiconflit*, comprising over 35 million tokens in total, as shown in Table 1. In parallel, a linguistic expert manually created rule-based automata to tag discourse markers, including connectives and enunciative particles (Dargnat, 2024). These rule-based automata, developed using the Unitex software [2], were applied to the corpora, automatically annotating the discursive and non-discursive usages of markers. When the discursive status of the marker could not be resolved by the automata, these ambiguous cases were labeled as candidate markers (CAND).

Since the automata were made by one expert and the resulting data were semi-automatically annotated, we deemed it crucial to validate the annotation quality against a gold-standard dataset. Therefore, we applied the automata to the raw text of the FDTB (Danlos et al., 2015) then compared the output to its annotated version.

Notably, the connective vocabularies of the FDTB and the automata are not identical as the FDTB focuses on connectives while the rule-based automata include enunciative particles. Despite this discrepancy, we observe a high agreement, achieving an F1-score of 0.76 when ambiguous connec-

[2] https://unitexgramlab.org/fr

	Evaluation			Support	
Annotation	**Prec.**	**Rec.**	**F1**	**FDTB**	**Corpus**
With CAND	0.67	0.87	0.74	759	1505
Without CAND	**0.83**	**0.72**	**0.76**	**759**	**439**

Table 2: **Automata Annotation Performance Analysis:** This table compares automata annotation quality based on the inclusion of candidate markers (CAND). Bold results highlight the optimal configuration where removing ambiguous connectives results in a higher F1-score.

tives are excluded, as described in Table 2. Furthermore, an in-depth analysis showed that 87.5% of the annotation disagreements occur on tokens that both the FDTB and the automata identify as connectives, with one system labeling the occurrence as discursive and the other as non-discursive. This illustrates that the cause of the errors lies in disambiguation rather than the vocabulary mismatch itself, proving that the vocabulary difference has a minimal impact on the evaluation.

Depending on the treatment of ambiguous connectives, the annotations reflect a trade-off between precision and recall. By excluding ambiguous annotations, the data becomes less noisy, leading to a superior overall balance as demonstrated by the F1 score. Since our objective is to train a model capable of generalizing the concept of discourse relations, we have opted to prioritize data that minimizes false positives.

Consequently, with an F1 score of 0.76 against manual annotations, we conclude that the automata-annotated dataset is suitable for training a model focused on discourse connectives. Despite being semi-automatically generated, the substantial scale of our annotated corpus provides a high volume of information that can be employed by BERT-based architectures.

3.2 The annotation of discourse relations

In order to train our model to classify the relation of a connective in a text, our training data requires each connective to be tagged with the appropriate relation. As there are no large-scale French corpora mapping explicit connectives to relations in context, we utilized LEXCONN, a lexicon of French connectives with the relations they can represent and some usage cases of these connectives.

A connective can be monosemous or polysemous. For example, in LEXCONN, *même si* (even if) is exclusively a connective of concession but *tandis que* (while) can express *Contrast* or *Background* depending on the context. Because of the volume of our training data, manually labeling the dataset proved to be challenging. Therefore, we only automatically tagged the relation of monosemous connectives, as each is associated with a single relation in the lexicon. Focusing on the annotation of connectives expressing one relation allows us to train our model on unambiguous data and the automatic labeling enriches the corpus with annotations of discourse relations.

By utilizing an annotated dataset of monosemous connectives and the contextually aware CamemBERT model, we hypothesized that fine-tuning would build robust internal representations of discourse relations. The Transformer-based architecture allows the model to learn these relations contextually from numerous examples of connectives during training. This capacity for generalization should, in theory, enable the model to perform classification on connectives not encountered during the training phase. Such an approach effectively addresses the scarcity of linguistic resources and specialized tools for French discourse analysis. In Section 4.3, we validate this hypothesis by evaluating the model on a manually annotated dataset featuring entirely unseen connectives.

LEXCONN indexes 23 discourse relations with monosemous connectives. However, the relations *Temploc* (Temporal Localisation), *Elaboration* and *Rephrasing* were removed from the training data because of insufficient data in terms of connectives or the number of training instances. Furthermore, *Opposition* and *Concession* revealed to be very similar semantically as they both express the violation of expectation between arguments. Consequently, we decided to merge the connectives of these two relations under *Concession*, which serves as a broader category. As a result, the automatically annotated corpus comprises 19 discourse relations.

Regarding context extraction, we retrieve the entire sentence containing the target connective. If the connective is located at the beginning of the sentence, we also retrieve the preceding sentence to ensure that the context window captures both the connective and its arguments. While we acknowledge that discourse unit segmentation would provide a more precise boundary for argument identification, we leave the integration of discourse unit segmentation into our pipeline for future work.

Table 8 in Appendix A.1 shows the distribution of the 19 discourse relations and how 189 con-

nective types are distributed across them in the automatically annotated corpus of connectives and relations.

3.3 An annotation protocol for an evaluation dataset

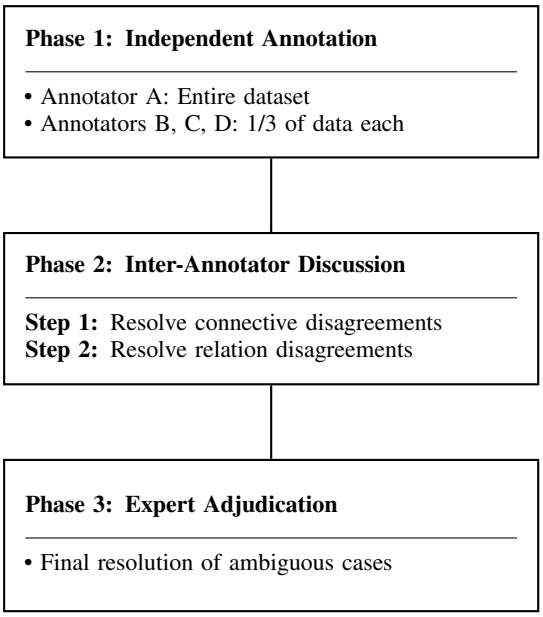

Figure 1: Three-phase annotation protocol for connective identification and discourse relation labeling.

Because of the absence of a French dataset annotating the relation of a connective, we created a curated dataset labeling connectives and their discourse relations to assess the performance of our model.

Data Collection : We identified monosemous connectives from LEXCONN that were not tagged in our annotated corpus of connectives and automatically extracted instances containing those connectives from the raw corpora presented in 3.1. We excluded polysemous connectives to simplify the annotation process as disambiguating their discourse relations is inherently more complex.

After collecting the sentences, a manual review revealed that some connectives were highly lexically ambiguous. We decided to remove them from the dataset to maximize the number of valid instances. This process resulted in a dataset of 331 instances.

Annotation Protocol : The frameworks of Danlos et al. (2015) and Muller et al. (2012) were used to design an annotation guide because they respectively propose a definition of connectives and SDRT relations in LEXCONN. We provided this guide to the annotators tasked with a two-step labeling process. Firstly, they had to indicate if the candidate connective is indeed a connective; if so, they were required to clarify if the given relation is the one expressed by the connective. Although the selected connectives from LEXCONN are monosemous, this protocol allowed us to verify this classification empirically and eventually identify exceptions.

	Connective			Relation		
Pair	κ	AC1	Agr. (%)	κ	AC1	Agr. (%)
A–B	0.341	0.636	71.171	0.226	0.771	80.000
A–C	0.331	0.580	67.592	0.178	0.791	81.356
A–D	0.385	0.656	77.678	0.288	0.818	83.784

Table 3: **Inter-Annotator Agreement (IAA) for Discourse Labeling:** This table summarizes agreement levels across three annotator pairs (A–B, A–C, and A–D) for two tasks: identifying the discourse **Connective** and classifying the **Relation**. Agreement is measured using Cohen's kappa (κ), Gwet's AC1, and raw percentage agreement (Agr. %).

Annotation Process : Figure 1 shows the annotation pipeline. One annotator processed the entire dataset while three other annotators analyzed each a portion of the data without overlap between them. Table 3 shows the inter-annotator agreement that was obtained by computing Cohen's Kappa coefficient, Gwet's AC1 agreement coefficient (Gwet, 2008) and the raw percentage agreement. For both connective and relation identification, κ values are lower than AC1. This occurs because our selection of monosemous connectives creates a heavy majority of positive instances in the data, which inevitably increases the probability of agreement. Cohen's κ heavily penalizes this agreement on majority classes by attributing it to chance. In contrast, Gwet's AC1 provides a more accurate representation of the actual agreement between annotators and aligns more closely with the observed average agreement.

The average agreement of 72% for connective identification reflects annotator A's stricter interpretation of the definition, leading them to favor non-connective labels, as detailed by the confusion matrices in Appendix A.2. On the contrary, the average agreement of 81% for the relation classification illustrates the robustness of LEXCONN in the identification of monosemous connectives but it also reveals borderline cases where connectives labeled as exclusively monosemous in LEXCONN proved to be polysemous during the manual anno-

tation process.

The independent annotation was followed by a two-phase meeting between the annotator pairs. First, they discussed the disagreement in the connective identification. This exchange allowed the analysis of candidate connectives initially negatively labeled; by consulting the extended context, annotators were able to definitively confirm or refute each instance's status. The relevant context was subsequently integrated into the source text, and labels were updated accordingly. Second, annotators assigned discourse relations to all instances reclassified as connectives. A follow-up discussion was held to resolve remaining disagreements regarding these relations. Finally, 42 unresolved cases were submitted to an expert for adjudication.

The expert adjudication confirmed the connective status and relation for 14 examples. However, 19 instances featured a connective expressing multiple relations at once; specifically, 3 examples featured connectives that appeared to express relations defined in SDRT and PDTB frameworks. Furthermore, in 4 cases, while polysemy was apparent, identifying the full set of expressed relations remained a challenge.

This protocol highlights the complexity of discourse relation identification, a task highly dependent on context. Nevertheless, the robust inter-annotator agreement observed for the relation labeling task shows the reliability of the LEXCONN lexicon across a wide range of connectives, despite linguistic variations. Notably, out of the 54 connectives in our annotated corpus, only 4 initially defined as monosemous in the lexicon proved to be empirically polysemous.

A total of 200 instances covering 54 connective types were manually annotated. The resulting dataset covers most discourse relations considered in our classification task, although a small number are not represented due to the limited availability of suitable connective instances. Table 4 details the distribution of instances by relation.

Relation	**Instances**	**Conn. Type**
Concession	24	9
Condition	20	6
Summary	18	5
Result	15	4
Explanation	14	3
Consequence	12	2
Goal	12	2
Contrast	12	3
Detachment	11	3
Continuation	10	3
Background	9	2
Alternation	9	2
Result*	8	3
Narration	8	2
Explanation*	7	3
Parallel	6	3
Commentary	3	1
Elaboration	2	2
Total	**200**	**54**

Table 4: **Detailed distribution of discourse relations in the manually annotated corpus** ($N = 200$): *Explanation** and *Result**, defined in SDRT, apply on speech acts and not on events unlike the other relations.

4 Experiments

4.1 Implementation details

Our model Relex is a CamemBERT-base model fine-tuned with the data that maps a connective in context to its discourse relation described in 3.2. The input of the model is a text sequence with a tagged connective and the model predicts a relation among 19. The tags are defined as special tokens in the model and the connectives can be a single word or multi-words as they are labeled directly in the text.

To overcome the unbalanced data between relations, we apply a weighted function that computes the inverse of the frequency of each relation, normalizes the results, and applies them to the loss function. The model was trained on 3 epochs with a learning rate of $2e^{-5}$, using the AdamW optimizer and a batch size of 16. The maximum sequence length was set to 256 tokens with dynamic padding.

4.2 Evaluation setup

Table 4 highlights an unbalanced distribution in the manually annotated dataset; specifically, there are no instances of Flashback and Evidence relations due to the lack of valid instances in the initial corpus. To address this and ensure a thorough assessment of Relex, the final evaluation dataset contains 269 instances covering 19 relations, combining manually annotated examples and 71 additional instances from LEXCONN to compensate for missing relations. Selected LEXCONN examples are usage cases of connectives unseen during the training phase. Moreover, as established by Roze et al. (2012), the examples within LEXCONN are extracted from the FRANTEXT [3] corpus. Therefore, their inclusion preserves the natural linguistic complexity of the evaluation set. The number of LEX-

[3] https://www.frantext.fr

Model	Params	Prec.	Rec.	F1	Acc.
Gemini 3 Flash	–*	**0.67**	**0.66**	**0.62**	**0.65**
Relex (CamemBERT-base)	110M	0.62	0.62	0.59	0.63
Relex (mBERT)	178M	0.42	0.40	0.38	0.41
CamemBERT + Log. Reg.	110M	0.28	0.21	0.21	0.22
Mistral 7B	7B	0.21	0.25	0.20	0.26

*Proprietary model; exact parameter count not officially disclosed.

Table 5: **Model performance comparison (Macro Average)** : While the generative Gemini 3 Flash achieves the highest scores, the fine-tuned CamemBERT-base (Relex) demonstrates superior efficiency, outperforming Mistral 7B and mBERT fine-tuned version.

CONN examples added to our manual annotated dataset is detailed in Table 12, Appendix A.3.

To the best of our knowledge, there is currently no French model specifically designed to predict the discourse relation expressed by a connective in context. Therefore, we trained a logistic regression classifier with the training data of Relex and used the frozen embeddings of CamemBERT. This classifier serves as a baseline because the embeddings were not fine-tuned for the task. We also used the open-source LLM Mistral 7B (Jiang et al., 2023) to perform the task prediction of Relex through few-shot prompting. The prompt includes one manually selected example per relation (19 in total) retrieved from the Relex training data. Those instances are chosen to provide one clear and representative instance for each discourse relation with the tagged connective. Predictions are generated with temperature set to 0 and constrained to the predefined set of relations using a JSON schema to prevent generation errors. The same prompting configuration and examples were used for the proprietary LLM Gemini 3 Flash to ensure a consistent comparison between models. Finally, as well as CamemBERT-base, we fine-tuned mBERT (Devlin et al., 2018) for the task.

4.3 Results

Table 5 reveals that Relex achieves notably higher performance than the classifier trained on frozen CamemBERT embeddings. The F1-score improvement of 0.38 underlines the importance of specializing word embeddings through fine-tuning, as those embeddings gain discursive knowledge thanks to the prediction task. Despite its strong general capabilities, Mistral 7B seems to fail at capturing the specificity of relations through connectives. Relex (mBERT) outperforms both the baseline and Mistral 7B. However, the multilingual model does not reach the performance level of CamemBERT, which is pre-trained on a French corpus. Finally, Relex demonstrates performance comparable to the latest version of Gemini Flash, Gemini 3 Flash.

Relation	Prec	Rec	F1	Supp.
Consequence	1.00	0.93	0.97	15
Goal	0.93	0.93	0.93	15
Summary	0.94	0.89	0.91	18
Background	0.93	0.87	0.90	15
Detachment	0.63	0.92	0.75	13
Contrast	0.83	0.67	0.74	15
Alternation	0.90	0.60	0.72	15
Explanation*	0.69	0.73	0.71	15
Commentary	0.80	0.57	0.67	7
Result	0.52	0.93	0.67	15
Concession	0.53	0.79	0.63	24
Narration	0.60	0.60	0.60	15
Condition	0.82	0.45	0.58	20
Parallel	0.70	0.47	0.56	15
Flashback	0.29	1.00	0.44	2
Explanation	0.35	0.40	0.38	15
Result*	0.25	0.07	0.11	15
Continuation	0.00	0.00	0.00	15
Evidence	0.00	0.00	0.00	5

Table 6: **Detailed performance of Relex by relation** : the connectives of the evaluation dataset are unseen during the training phase.

Table 6 shows that Relex achieves an F1-score above 0.5 for 14 relations. However, for the *Evidence* and *Flashback* relations, the limited support in the test set is insufficient to definitively confirm the model's ability to generalize.

Confusion between similar relations: Relex tends to mistake *Evidence* and *Flashback* with semantically similar relations. The confusion matrix of the model (see Figure 3 in A.3) reveals that out of 5 false negative cases for *Evidence*, Relex predicts *Result* in three cases; both relations belong to the causal category described by (Danlos et al., 2012).

Likewise, three examples of *Narration* are predicted as *Flashback* which is the inverse of *Narration*, as it describes the same relation but with reverse temporality. The model also fails to predict *Continuation* and six instances of this relation are

labeled *Narration*. *Narration* is a relation depicting situations in the order when they occur. Therefore, it encodes a succession in actions that can also appear with the expression of *Continuation* because the argument attached to the connective is the continuation of the first argument and adds supplementary information. The sequentiality encoded in both the relations may explain the mistakes of the model.

Systematic misclassification of specific connectives: Furthermore, all of those instances of confusion between *Continuation* and *Narration* involve the connective *"et puis"* (and then) and a pattern that emerges with Relex is a tendency to systematically map a connective to an incorrect relation. For instance, the model assigns the *Explanation* connective *"par le fait que"* to *Continuation* or the *Condition* connectives *"a fortiori si"* and *"surtout si"* to *Explanation*. However, those relations belong to different categories : *Explanation* represents causal effects, *Continuation* introduces additive information and *Condition* is a logical relation.

Compositionality of multi-word connectives: Moreover, it is interesting to observe that this misclassification of some connectives involves multi-word connectives, which can sometimes have a compositional dimension. For example, while the connective *"si"* almost unanimously marks a condition, the presence of terms such as *"a fortiori"* and *"surtout"* seems to reinforce the argument introduced by the connective, which could be semantically perceived as a justification or explanation. To further study this hypothesis and apprehend the behavior of the model, we will focus on a polysemous connective *mais* (but) in Section 4.4.

Even though Gemini 3 Flash obtains a higher F1 for the relations where Relex has an F1 below 0.5 (see Table 13 in Appendix A.3), the results of the LLM also remain below 0.5.

Connectives remain complex linguistic cues and their meanings are influenced by both their lexical forms and the surrounding context. Overall, Relex achieves a macro F1 score of 0.59 on a challenging 19-class discourse relation classification task.

4.4 A study of *mais*, a polysemous connective

To better understand Relex, we studied its predictions for polysemous connectives. From the automatic corpus described in 3.1, we extracted 187,805 instances containing 39 polysemous connectives from LEXCONN. Each connective can express 2 to 3 relations. As explained in 3.2, *Opposition* and *Concession* were merged into *Concession*.

We focused on *"mais"* (but), the most frequent polysemous connective in the dataset. The choice of a general analysis of a connective rather than an evaluation of the model on polysemous connectives reflects the lack of a French benchmark, making a full evaluation of all polysemous connectives impractical.

"Mais" can express *Concession* or *Contrast*, both part of the comparative category described by Danlos et al. (2012). Relex predicts 10 relations for the 119,867 instances containing *"mais"*, as detailed in Table 7.

Relation	Instance	Manual Analysis
Concession	119,745	20
Continuation	58	5
Narration	21	5
Explanation*	15	5
Explanation	14	5
Detachment	5	5
Commentary	5	5
Flashback	2	2
Contrast	1	1
Condition	1	1
Total	119,867	50

Table 7: Relation predictions of *"mais"* by Relex and the distribution of manual analysis by relation.

For manual analysis, we sampled 50 instances across the 10 predicted relations. Table 7 reports the distribution. We also applied SHAP (Zhao et al., 2024) to estimate the impact of each word on Relex predictions. These two approaches guided our analysis.

Table 7 reveals that Relex mainly predicts *Concession*, one of the expected relations. This indicates that the model uses both knowledge and context, as the specific mapping of *"mais"* was not seen during training. However, Relex predicts *Contrast* only 1 time. We initially hypothesized that the higher number of *Concession* instances in the training data might bias the model, despite the weighted loss function designed to alleviate this imbalance. However, the connective *somme toute* (all in all) contradicts this possibility: although Concession training instances outnumber *Summary*, Relex predicts *Summary* 63 times and *Concession* 56 times. This shows that the model is not systematically biased by training distribution. Because *Contrast* and *Concession* both highlight a comparison between two events, their semantic proximity could pose a

significant challenge for the model, likely causing this discrepancy in the predictions.

Other relation predictions occur when *"mais"* is combined with another connective. In these cases, the compositional meaning appears mainly influenced by the additional connective, affecting the predicted relation. In Figure 2, SHAP values highlight the impact of *"étant donné que"* on the prediction of *Explanation* for Example (2).

(2) **FR** : *Je reste favorable pour des raisons méthodologiques* ***mais*** *étant donné que je ne pense pas que cela induise quoi que ce soit du point de vue du sens, je m'en remets à vos opinions.* → **Prediction : Explanation**
(*English* : *I remain in favor of this for methodological reasons,* ***but*** *since I don't think it changes anything in terms of meaning, I'll leave it up to you.*)

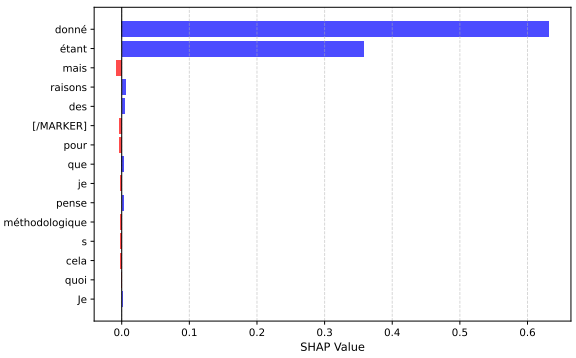

Figure 2: SHAP values highlighting the impact of the connective *étant donné que* in the prediction of *Explanation* ; the top 15 words are displayed.

"Mais" can also appear with connectives absent from training, such as *"ensuite"* (then), a polysemous connective expressing *Narration* or *Continuation*. Even when a connective is unseen, Relex can often infer its relation from context. Table 14 in Appendix A.4 lists connectives combined with *"mais"*, present or not in the training data, when the predicted relation differs from *Concession*.

Finally, our sample analysis and SHAP results reveal that certain syntactic structures also influence predictions. For example, the pattern *pas/non seulement ... mais aussi/surtout* (not only ... but also) occurs in 26 of 48 *Continuation* predictions. In these cases, *"mais"* is part of an enumerative structure, so the model predicts *Continuation* rather than *Concession* or *Contrast*. Similarly, 7 of 15 *Explanation** predictions involve *"mais"* in interrogative arguments, suggesting that the model sometimes relies on syntactic cues over semantic ones.

In summary, this manual study highlights how connective combination and syntactic patterns can shift Relex predictions away from the expected relation. While these observations provide a compelling explanation for the model's behavior, they represent a preliminary hypothesis derived from a manually inspected subset.

5 Discussion

Creation of discourse datasets: In this work, we developed two French datasets that enrich discursive resources and can be used for NLP tasks.

First, we evaluated a semi-annotated corpus of discourse markers against a manually annotated dataset. With an F1 of 0.76, we validated its relevance for a training task. We also automatically added discourse relations tags, producing a dataset used to fine-tune Relex, a CamemBERT-based model for predicting the relation expressed by a connective in context. One potential improvement of the quality of the dataset could be to segment the data in discourse units rather than sentences, as discourse units better capture the full scope of a connective's arguments. DISRPT project (Braud et al., 2025) provides a French discourse unit segmentation model, which could be integrated into our pipeline in future work.

Second, we collected a manually annotated dataset of 200 instances covering 18 LEXCONN relations. This dataset, labeled by a pair of annotators and reviewed by an expert for edge cases, provides real-world data to evaluate automatic models. It is released under CC BY-NC-SA 4.0.

Annotation model for discourse relations of connectives in context: Relex is a CamemBERT-based model that acts as a bridge between a connective in context and the relation it expresses. Trained on monosemous connectives, Relex can be applied to unseen connectives thanks to the contextually-aware architecture of BERT. Unlike static lexicons which only enumerate potential meanings, Relex identifies the relation of a connective in context. Consequently, it can serve as a first-pass automatic annotator, able to shift the human workload from manual annotation to verification. In addition, Relex is computationally inexpensive compared to massive language models with billions of pa-

rameters. Moreover, despite its smaller size, its performance is comparable to large models such as Gemini 3 Flash. Finally, as an open-source model, it is reproducible and accessible to the community.

6 Conclusion

This work presents the creation of two annotated French discursive datasets and a fine-tuned model for predicting the discourse relation of a connective in context. A semi-automatically annotated corpus for connectives enriched with relations from LEXCONN was leveraged to fine-tune a CamemBERT-base architecture. The resulting model, Relex was trained to classify a connective into one of 19 discourse relations, based on its specific context. Evaluated on a manually annotated dataset comprising connectives unseen during training, Relex achieves a macro F1-score of 0.59.

To better understand the model's behavior, we also conducted a qualitative analysis focused on the highly frequent polysemous connective *"mais"*. A manual study of a data sample, supported by SHAP values, revealed that Relex's predictions are highly sensitive to the meaning of combination of connectives and specific syntactic structures, which can sometimes override purely semantic cues. Future work will aim to validate these hypotheses on broader datasets and explore methods to mitigate potential syntactic biases.

7 Limitations

While this work introduces novel discursive datasets and a relation prediction model for French, several limitations must be acknowledged.

First, regarding the data, our manually annotated corpus remains limited in size (200 instances). While it constitutes a first resource for connective-to-relation mapping in French, expanding this corpus is necessary to have a representative amount of data for each relation expressed in LEXCONN. Furthermore, our quantitative evaluation is currently constrained to monosemous connectives. As noted during our qualitative analysis, the lack of an existing French benchmark makes a full quantitative evaluation of polysemous connectives challenging at this stage. Creating a balanced, gold-standard dataset dedicated to polysemous disambiguation represents a critical next step for future work.

Second, the proposed model architecture presents two main constraints. The model assumes the prior identification of connectives and therefore does not perform connective detection. As a result, external pipelines must be used for connective identification or disambiguation in French (Laali and Kosseim, 2017a). In addition, the training setup is restricted to the 19 LEXCONN relations considered in this work. However, exploring other theoretical discourse frameworks would require the creation of additional training resources.

Finally, regarding the error analysis of Relex, our investigation on connective combinations (e.g., *mais surtout*) and syntactic structures (e.g., *non seulement ... mais aussi*) was conducted on a manually analyzed sample. While SHAP token-attribution confirmed the mechanism of relational overriding in these cases, a broader automated analysis of the full dataset is required to confirm the statistical prevalence of this theory. We leave this large-scale validation for future work.

Acknowledgements

This work was financially supported by the CODIM project (ANR-22-CE38-0002). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.[4] Additionally, generative AI was utilized for coding assistance but all algorithmic logic and prompts were designed by the authors and the resulting code was manually verified and refined.

The authors would also like to thank the annotators who participated in the creation of the manually annotated dataset.

References

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025), pages 1–20, Suzhou, China. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Wong, and Simon See. 2023. Self-consistent narrative prompts on abductive natural language inference. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association

[4] https://www.grid5000.fr

for Computational Linguistics (Volume 1: Long Papers), pages 1040–1057.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le FDTB: French discourse tree bank (towards the FDTB: French discourse tree bank)[in french]. In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN, pages 471–478.

Laurence Danlos, Margot Colinet, and Jacques Steinlin. 2015. FDTB1, première étape du projet «French discourse treebank»: repérage des connecteurs de discours en corpus. Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, (17).

Mathilde Dargnat. 2024. Les particules énonciatives. Encyclopédie Grammaticale du Français.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1):29–48.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

Majid Laali and Leila Kosseim. 2017a. Automatic disambiguation of french discourse connectives. Int. J. Comput. Linguistics Appl., 7:11–30.

Majid Laali and Leila Kosseim. 2017b. Automatic mapping of French discourse connectives to PDTB discourse relations. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pages 1–6, Saarbrücken, Germany. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure, volume 3, pages 87–124.

William Mann, Christian Matthiessen, and Sanda Thompson. 1989. Rhetorical structure theory and text analysis. Discourse Description: Diverse Linguistic Analyses of a Fund Raising Text, page 66.

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, pages 7203–7219, Online. Association for Computational Linguistics.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6277–6295.

Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Draoulec, and Laure Vieu. 2012. Manuel d'annotation en relations de discours du projet annodis. .

Marie-Paule Péry-Woodley, Stergos D Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher. 2011. Le corpus annodis, un corpus enrichi d'annotations discursives [the annodis corpus, a corpus enriched with discourse annotations]. Traitement Automatique des Langues, 52(3):71–101.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. Lexconn: a french lexicon of discourse connectives. Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, (10).

Manfred Stede and Carla Umbach. 1998. Dimlex: A lexicon of discourse markers for text generation and understanding. In COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague discourse treebank 3.0. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1270–1279, Torino, Italia. ELRA and ICCL.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5908–5923.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task, pages 1–16.

Deniz Zeyrek and Murathan Kurfalı. 2017. Tdb 1.1: Extensions on turkish discourse bank. In Proceedings of the 11th Linguistic Annotation Workshop, pages 76–81.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2):1–38.

A Appendix

A.1 Training Data

The monosemous connectives of LEXCONN annotated in the French corpora (see Table 1) are mapped to the relations defined in LEXCONN. Table 8 details the distribution of instances by relation in the automatically annotated corpus.

Relation	Instances	Conn. Type
Goal	120,185	27
Concession	43,162	38
Narration	23,002	3
Explanation*	17,735	12
Continuation	11,362	14
Result	6,152	29
Contrast	3,137	3
Flashback	2,965	4
Explanation	2,595	12
Background	2,019	5
Condition	1,462	19
Result*	1,436	2
Alternation	1,235	6
Detachment	1,139	6
Evidence	1,098	2
Consequence	846	2
Summary	265	2
Commentary	126	2
Parallel	88	1
Total	240,009	189

Table 8: **Distribution of discourse relations in the automatically annotated corpus** : *Explanation** and *Result**, defined in SDRT, apply on speech acts and not on events unlike the other relations.

A.2 Manual Annotation and Evaluation Dataset

Tables 9, 10, and 11 present the confusion matrices for each pair of annotators. They show the distribution of annotations for (a) connective identification and (b) relation identification.

Annotator Pair A–B

		Ann. B		
		Yes	No	Ambig
Ann. A	Yes	64	3	0
	No	24	15	0
	Ambig	5	0	0

(a) Connective Identification

		Ann. B		
		Yes	No	Ambig
Ann. A	Yes	49	10	0
	No	2	3	0
	Ambig	1	0	0

(b) Relation Identification

Table 9: Confusion matrices for pair A–B.

Annotator Pair A–C

		Ann. C		
		Yes	No	Ambig
Ann. A	Yes	57	3	1
	No	27	16	4
	Ambig	0	0	0

(a) Connective Identification

		Ann. C		
		Yes	No	Ambig
Ann. A	Yes	47	6	4
	No	0	1	0
	Ambig	0	1	0

(b) Relation Identification

Table 10: Confusion matrices for pair A–C.

Annotator Pair A–D

		Ann. D		
		Yes	No	Ambig
Ann. A	Yes	74	4	0
	No	21	13	0
	Ambig	0	0	0

(a) Connective Identification

		Ann. D		
		Yes	No	Ambig
Ann. A	Yes	60	10	0
	No	0	2	0
	Ambig	0	2	0

(b) Relation Identification

Table 11: Confusion matrices for pair A–D.

A.3 Model Evaluation

Table 12 shows the distribution of the dataset used to evaluate Relex.

Relation	Manual	Lexconn	Total	Conn
Concession	24	-	24	9
Condition	20	-	20	6
Summary	18	-	18	5
Explanation	14	1	15	4
Result	15	-	15	4
Alternation	9	6	15	5
Parallel	6	9	15	9
Background	9	6	15	7
Goal	12	3	15	5
Consequence	12	3	15	3
Result*	8	7	15	5
Narration	8	7	15	7
Continuation	10	5	15	6
Explanation*	7	8	15	9
Contrast	12	3	15	6
Detachment	11	2	13	5
Commentary	3	4	7	4
Evidence	-	5	5	5
Flashback	-	2	2	2
Total	**198**	**71**	**269**	**106**

Table 12: **Distribution of the evaluation dataset:** Examples from LEXCONN are added to the manually annotated dataset to ensure that each relation can be evaluated.

A.3.1 Relex

Figure 3 shows the confusion matrix of Relex (CamemBERT-base).

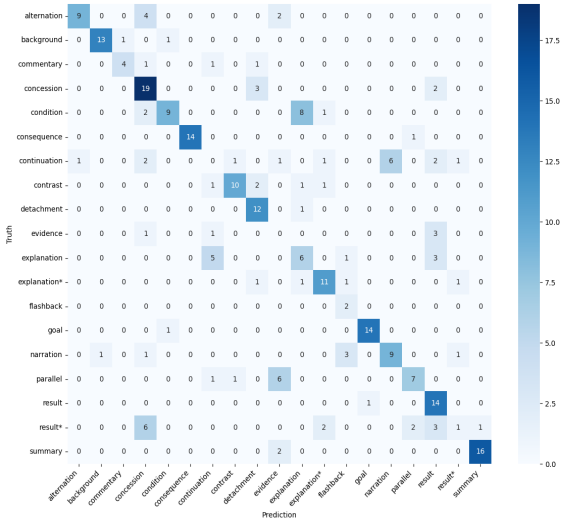

Figure 3: Relex evaluation confusion matrix

A.3.2 Gemini 3 Flash

Figure 4 shows the confusion matrix of Gemini 3 Flash.

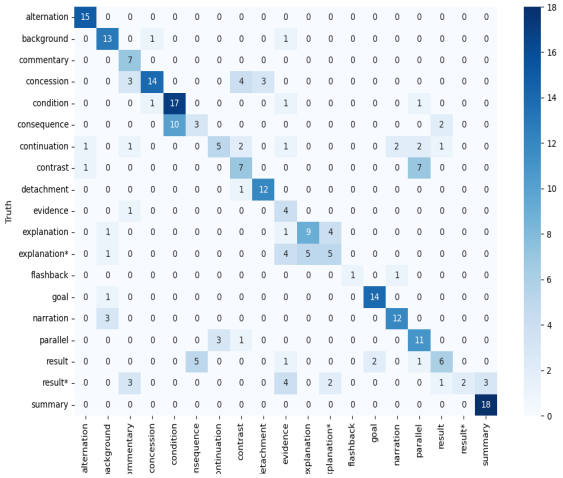

Figure 4: Gemini 3 Flash evaluation confusion matrix.

Table 13 shows the classification report by relation of Gemini 3 Flash on the evaluation dataset.

Relation	Prec	Rec	F1	Supp.
Alternation	0.88	1.00	0.94	15
Summary	0.86	1.00	0.92	18
Goal	0.88	0.93	0.90	15
Detachment	0.80	0.92	0.86	13
Narration	0.80	0.80	0.80	15
Background	0.68	0.87	0.76	15
Condition	0.63	0.85	0.72	20
Concession	0.88	0.58	0.70	24
Flashback	1.00	0.50	0.67	2
Commentary	0.47	1.00	0.64	7
Explanation	0.64	0.60	0.62	15
Parallel	0.50	0.73	0.59	15
Result	0.60	0.40	0.48	15
Contrast	0.47	0.47	0.47	15
Continuation	0.62	0.33	0.43	15
Explanation*	0.45	0.33	0.38	15
Evidence	0.24	0.80	0.36	5
Consequence	0.38	0.20	0.26	15
Result*	1.00	0.13	0.24	15

Table 13: Detailed performance of Gemini 3 Flash by relation.

A.4 Qualitative analysis of *"mais"*

Table 14 shows the connectives combined to *"mais"* and which likely triggered the prediction of Relex.

Relation	Train Conn.	New Conn.
Condition	à condition que (on condition that)	–
Detachment	de toute façon (anyway), de toute manière (in any case)	–
Flashback	une fois que (once), auparavant (previously)	–
Explanation	étant donné que (given that)	surtout que (especially since)
Narration	avant de (before)	ensuite (then)
Continuation	outre que (besides)	surtout (especially)
Contrast	–	au contraire (on the contrary)
Explanation*	–	puisque (since), parce que (because)

Table 14: Connectives appearing in combination with *mais* and their predicted relations, with English translations provided in parentheses.