

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

Fatou Sow^{1,2}, Yannick Toussaint¹, Mathieu Constant²

¹LORIA, Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

²ATILF, Université de Lorraine, CNRS, F-54000 Nancy, France

The Challenge: Resources for French Discourse Connectives

Speakers produce **coherent** utterances by using **connectives** such as *because*, *while*, so to signal **discourse relations**.

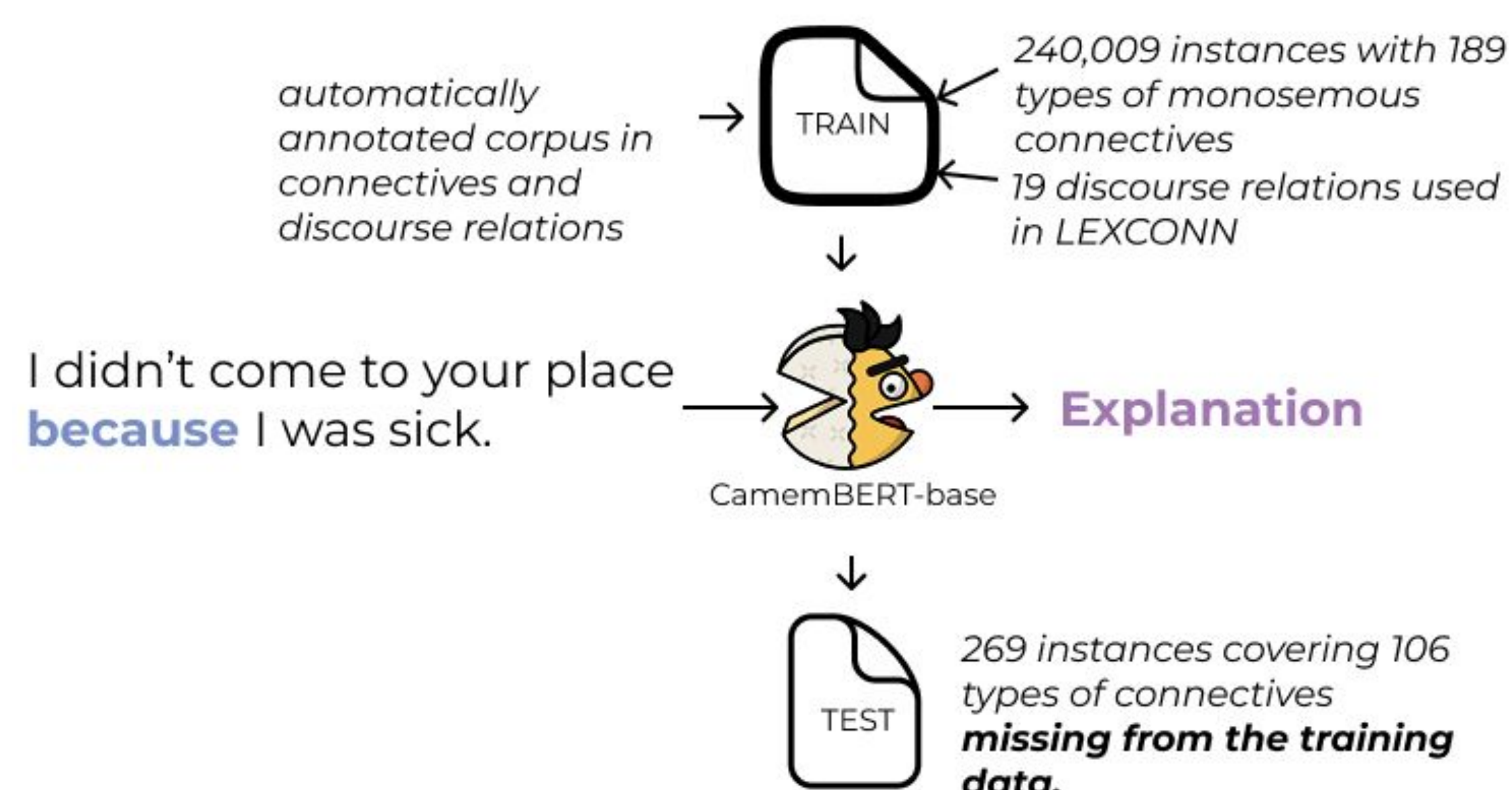
I didn't come to your place **because** I was sick.

Oh no ! Are you feeling any better now ?

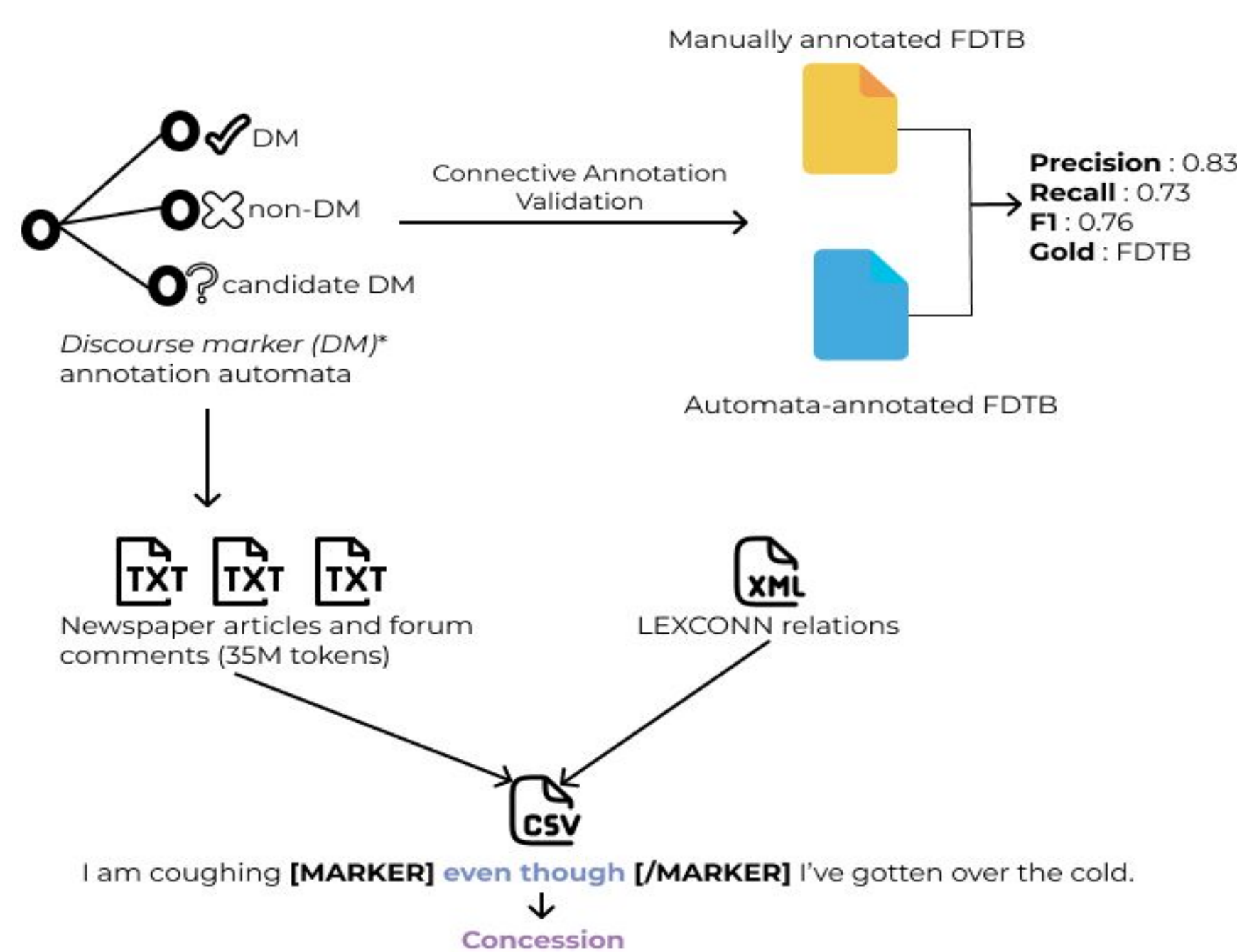
Yes.. Well, I am still coughing **even though** I've gotten over the cold.

- One important aspect of studying **discourse connectives** concerns the **resources** needed to train ML models. Such resources are **more developed** for some languages, like **English**, whereas **French** has far **less data**. The main available resources in French are the FDTB [1] and Annodis [2] corpora, as well as the LEXCONN lexicon [3].
- To the best of our knowledge, there is no existing French corpus annotating **both** connectives and SDRT discourse relations, nor a model designed to automatically annotate the relation of a connective in context.

Contributions: Novel Datasets and a Fine-Tuned BERT Classifier



An Automatically Annotated Dataset of Connectives and Relations



- Annotating only monosemous **connectives** ensures unambiguous training data while automatically enriching the corpus with discourse **relation** labels.
- This data volume enables model **training** for **connective**-related tasks.

Building the Evaluation Set



Pair	Connective			Relation		
	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

We annotated 200 instances of **connectives** in context and added 71 examples from LEXCONN. Relex was evaluated on these 269 instances, covering all **relations** in the training set.

Relex: Model Architecture and Results

We introduce **Relex**, a CamemBERT-base model fine-tuned for a 1-vs-19 **discourse relation** classification of **connectives** in context. Training the model on monosemous **connectives** provides unambiguous data, while the large dataset volume exposes the model to numerous contextual examples of each **relation** type. This enables Relex to build rich internal representations of discourse relations and apply this learned knowledge to **connectives never encountered** during training.

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.

While Gemini 3 Flash achieves the highest scores, Relex (110M params) remains highly effective and easily outperforms larger models like Mistral 7B.

Conclusion and Future Work

We created dedicated resources to train and evaluate models on **discourse connectives** and we presented a model that predicts contextual **relations** expressed by unseen **connectives**. Future work includes integrating **connective** detection into the Relex pipeline and expanding our evaluation dataset to cover polysemous **connectives**.

Bibliography

- [1] Danlos, L., Antolinos-Basso, D., Braud, C., & Roze, C. (2012, June). 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* (pp. 471-478).
- [2] 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.
- [3] Charlotte Roze, Danlos Laurence, Philippe Muller. LEXCONN: a French Lexicon of Discourse Connectives. *MAD 2010 - 8th Workshop Multidisciplinary Approaches to Discourse*, Mar 2010, Moissac, France. pp.114-125. (inria-00511615)

Datasets & Code

