Discourse Marker Identification in French Spoken Corpora: Using Rule-Based Method and Machine Learning

Presented by
Dahou Abdelhalim Hafedh

Supervised by
Mathilde Dargnat, Jacques Jayez,
Mathieu Constant

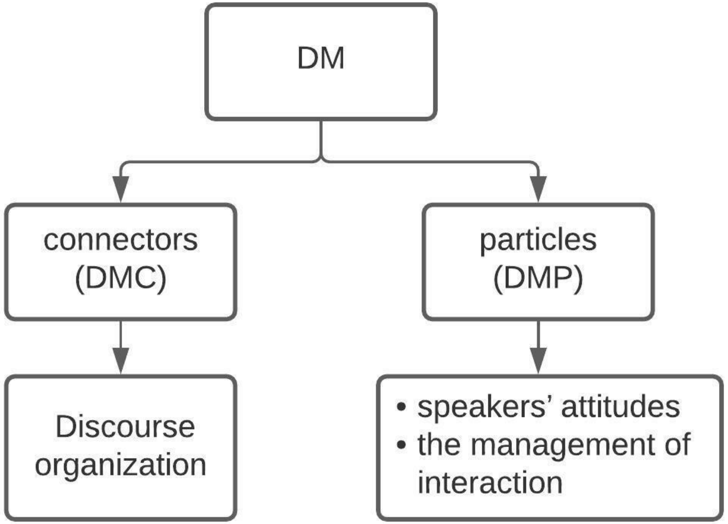
September 7th, 2021

Plan

1. Introduction

DM are linguistic expressions that have been proven to be effective for :

- Segmenting discourse into meaningful units.
- Recognizing relationships between these units.



2. Problem

(1) Nous avons passé/sommes restés un ***bon*** moment chez nos voisins.

We stayed with a good/long time at our neighbors.

(2) A – je vais te faire un super cadeau pour ta fête.

B – ***bon*** j'ai hâte de voir ça

A – I'll give you a great gift for your party.

B – DM, I cannot wait to see that

3. Contribution

1. We built four mechanisms based on rule-based and machine learning.
2. We tested different hypothesis and applied several scenarios.
3. Made experiment for each mechanism on the same spoken corpora.
4. We explored and evaluated the UNITEX platform.

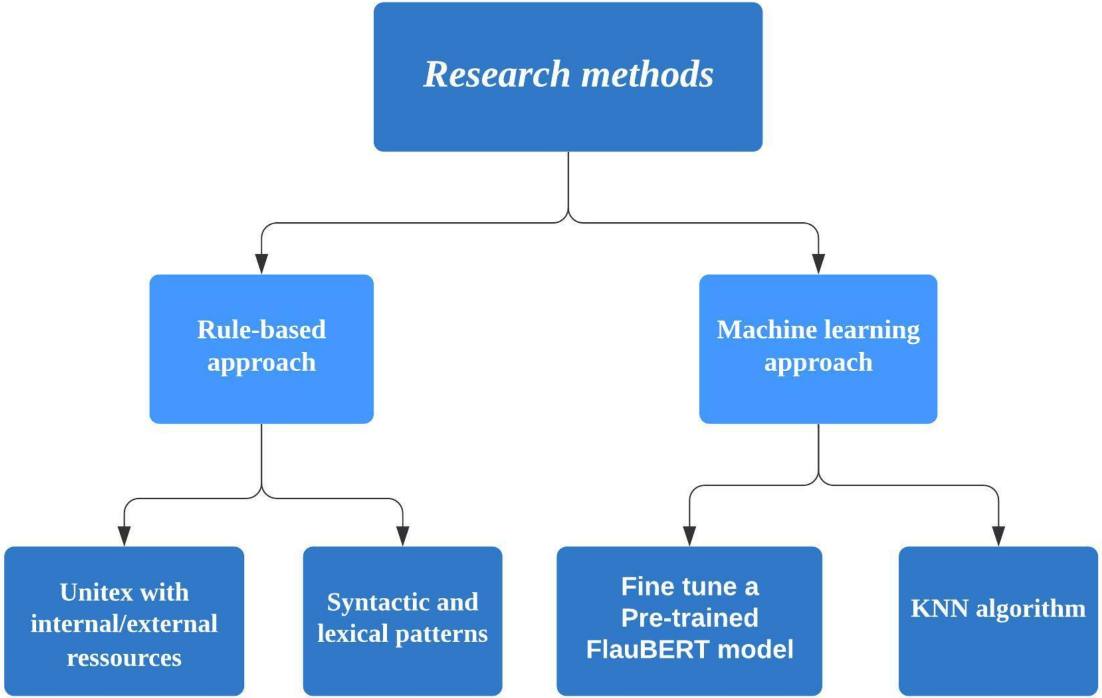
4. Data

1. **CORPAIX** is a corpus of spoken French that contains 941624 token, 40241 segment and 33 speaker.
2. **ESLO** is a concatenation of two French spoken corpus (ESLO 1 and ESLO 2) which contains 649081 token, 53772 segment and 28 speaker.
3. **TCOF** is a French spoken corpus of size 149292 token and 19527 segment.

Corpus	Attention	bon	la preuve	quoi
CORPAIX	171	3867	8	2380
ESLO	163	1797	11	958
TCOF	18	545	0	862

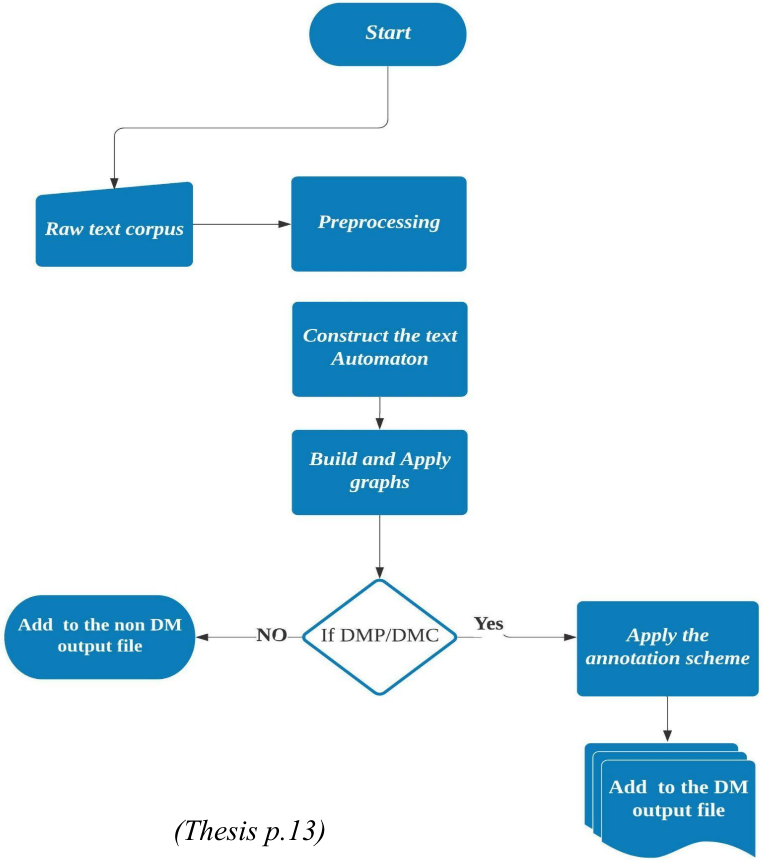
Table 1 : Distribution of DM in the spoken corpus.

5. Methodology



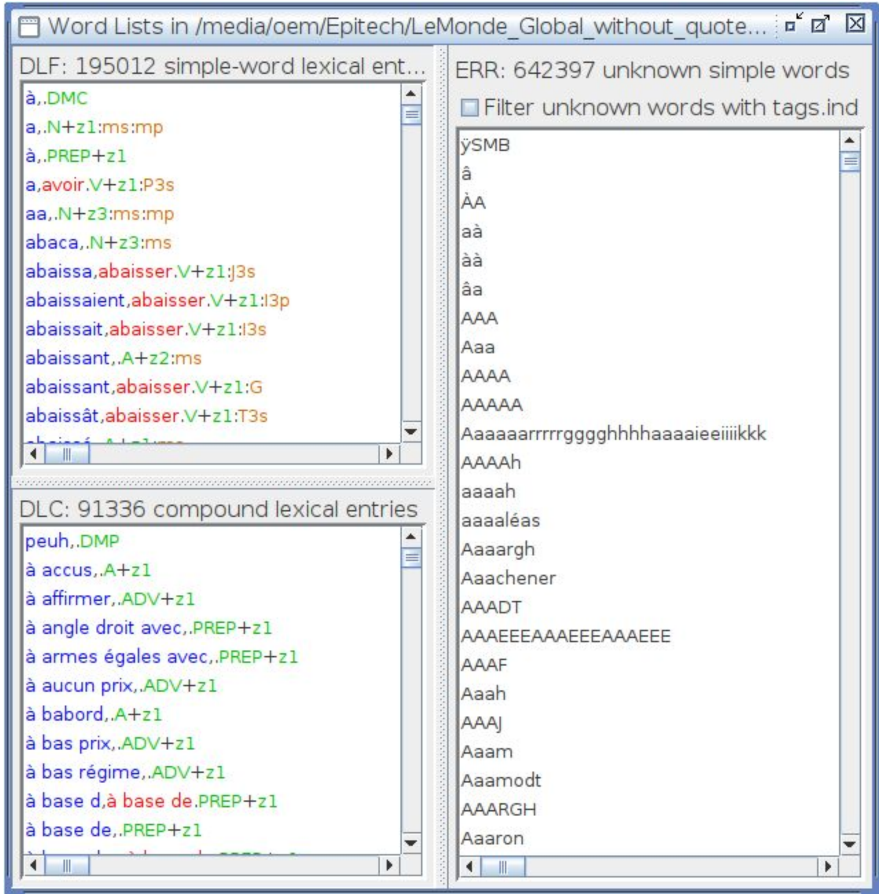
5. Methodology

5.1.a Unitex with internal resources



(Thesis p.13)

1. Read and preprocess the spoken corpus.
2. Construct the text automaton.
3. Build and apply graphs.
4. Visualization of annotation.



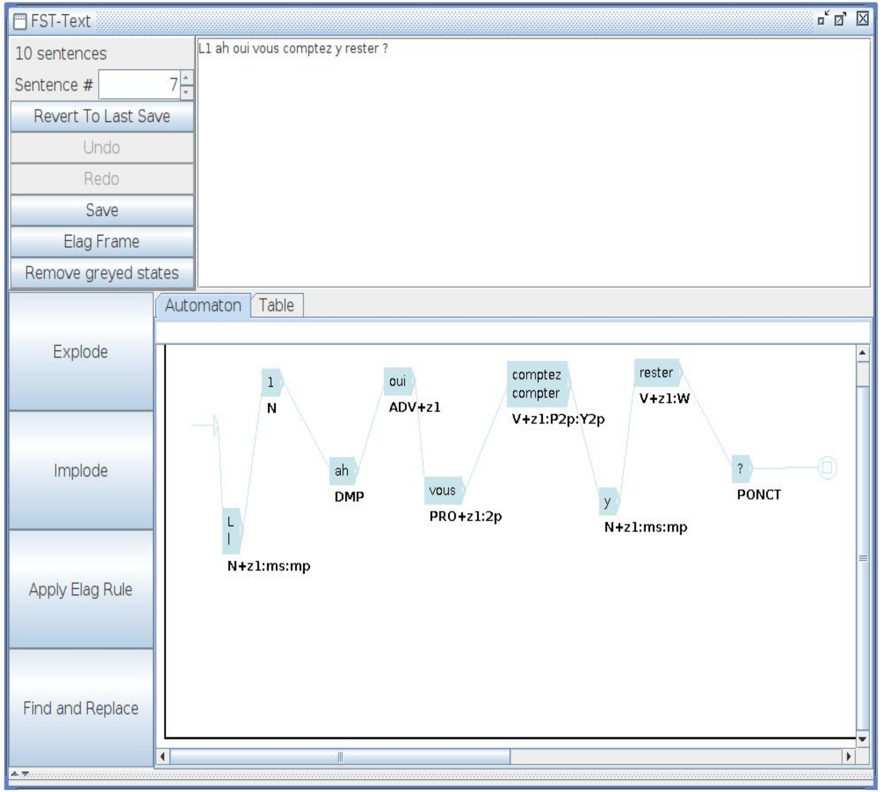
5. Methodology

5.1.a Unitex with internal resources

1. Read and preprocess the spoken corpus.
2. Construct the text automaton.
3. Build and apply graphs.
4. Visualization of annotation.

5. Methodology

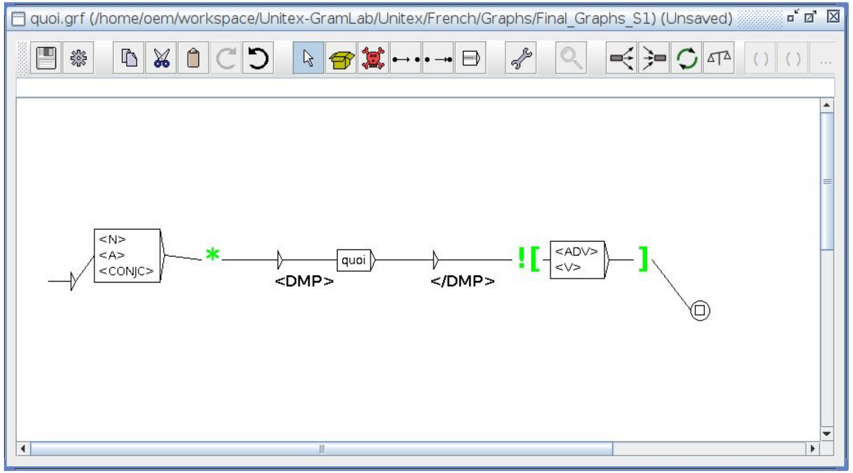
5.1.a Unitex with internal resources



1. Read and preprocess the spoken corpus.
2. Construct the text automaton.
3. Build and apply graphs.
4. Visualization of annotation.

5. Methodology

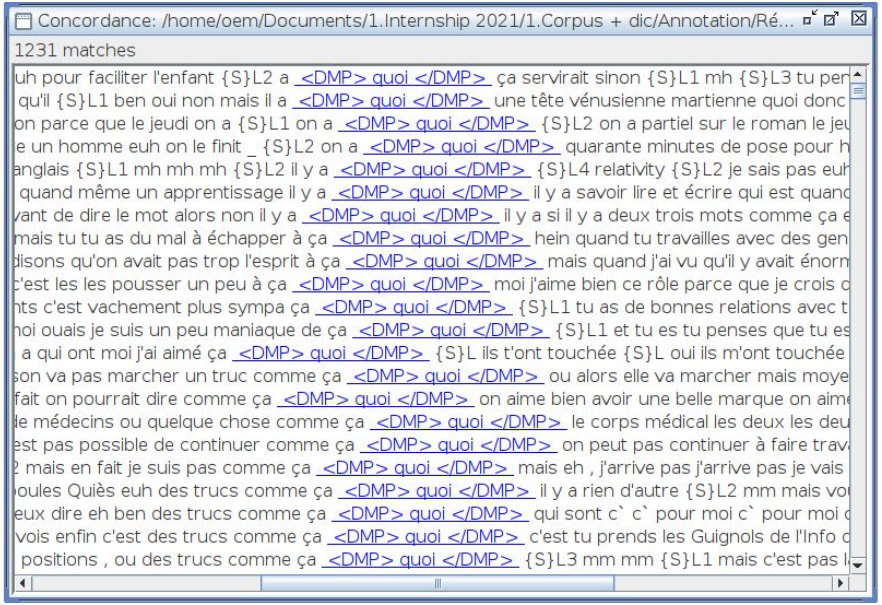
5.1.a Unitex with internal resources



1. Read and preprocess the spoken corpus.
2. Construct the text automaton.
3. Build and apply graphs.
4. Visualization of annotation.

5. Methodology

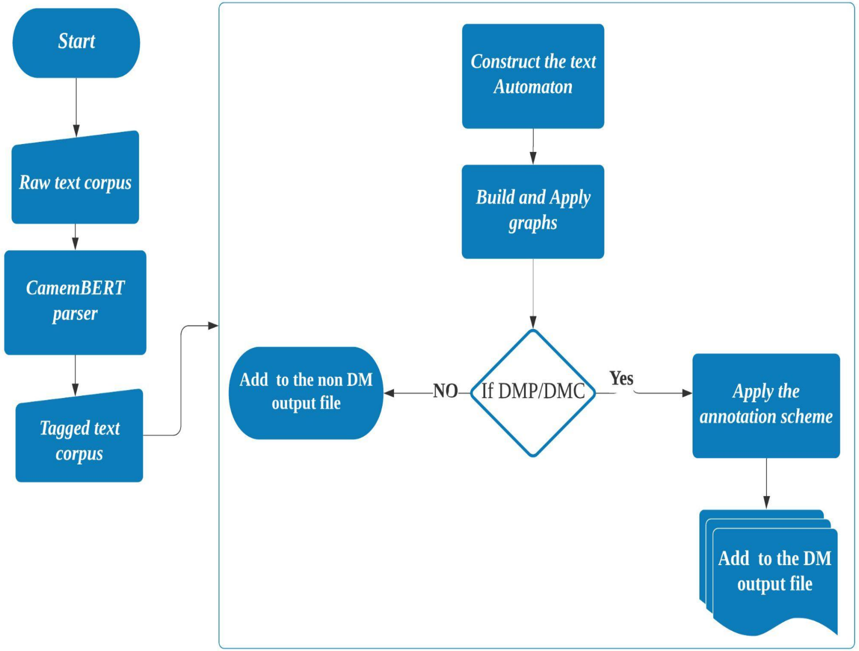
5.1.a Unitex with internal resources



1. Read and preprocess the spoken corpus.
2. Construct the text automaton.
3. Build and apply graphs.
4. Visualization of annotation.

5. Methodology

5.1.b Unitex with external resources

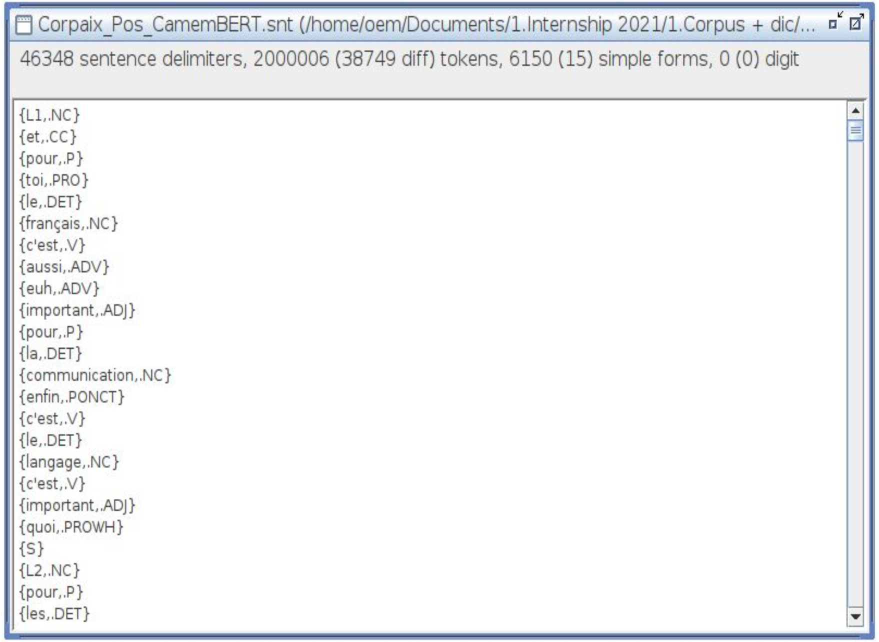


(Thesis p.20)

1. Tagged text with CamemBERT.
2. Unitex reading tagged corpus.
3. Build and apply graphs.

5. Methodology

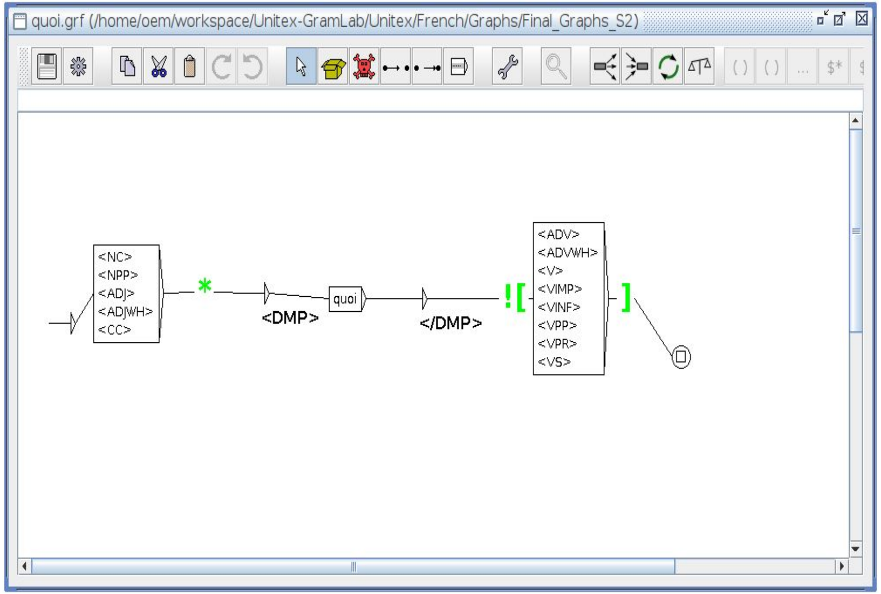
5.1.b Unitex with external resources



1. Tagged text with CamemBERT.
2. Unitex reading tagged corpus.
3. Build and apply graphs.

5 Methodology

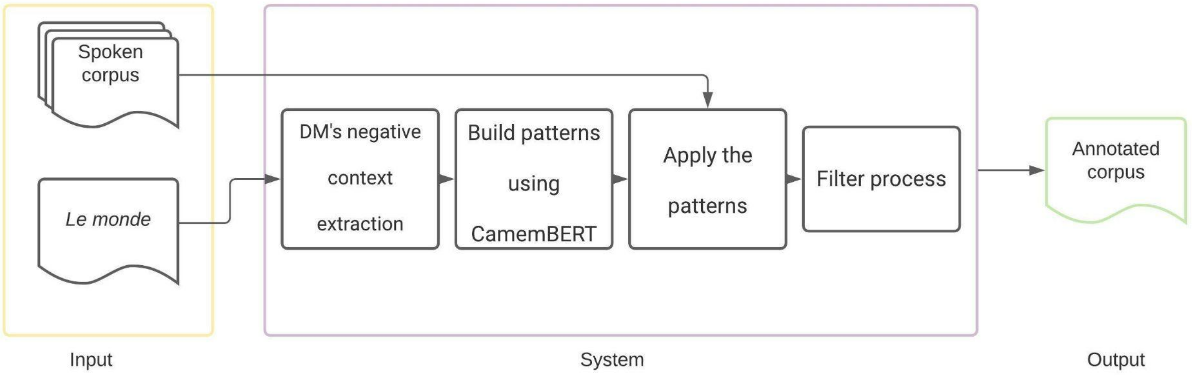
5.1.b Unitex with external resources

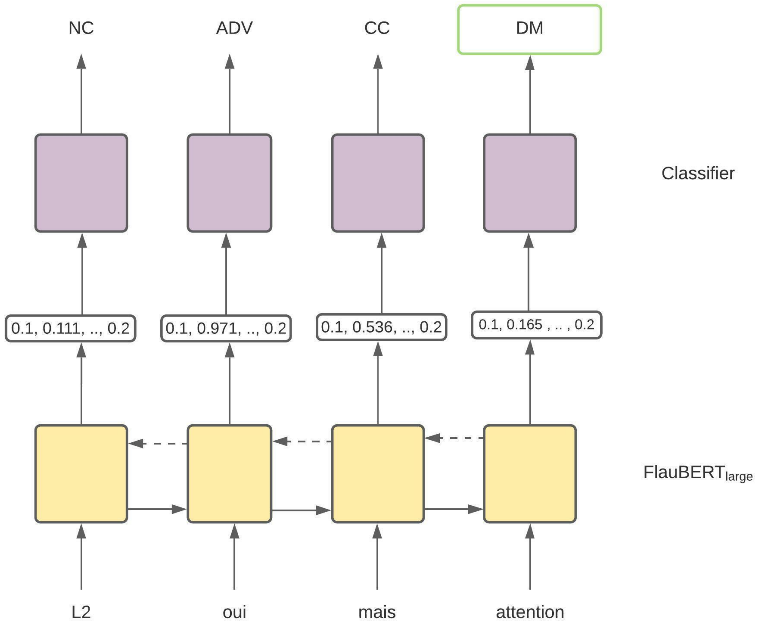


1. Tagged text with CamemBERT.
2. Unitex reading tagged corpus.
3. Build and apply graphs.

5. Methodology

5.2. Syntactic and lexical patterns





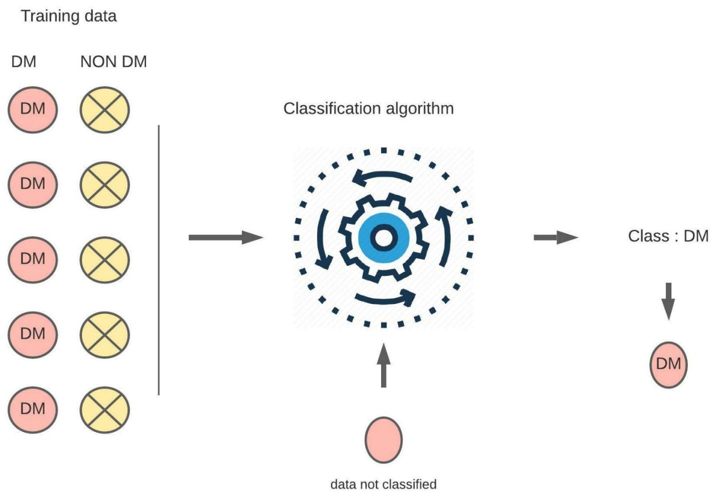
5. Methodology

5.3. Fine-tuning Pretrained FlauBERT model for PoS Tagging task

1. Data preparation
2. Building model.
3. Train and test the model.

5. Methodology

5.4. KNN algorithm with FlauBERT word embedding



1. Using a labeled dataset for training.
2. Fixing the K value.
3. Classification based on Cosine distance.

6. Evaluation and discussion

Unitex with internal and external resources

Corpus	Precision	Recall	F1-score
CORPAIX	0.91	0.49	0.63

Table 1 : Evaluation of unitex with its internal resources on the identification of marker ***bon*** in 100 samples from the ***CORPAIX*** corpus .

Corpus	Precision	Recall	F1-score
CORPAIX	1	0.49	0.65

Table 2 : Evaluation of unitex with external resources on the identification of marker ***bon*** in 100 samples from the ***CORPAIX*** corpus.

6. Evaluation and discussion

Syntactic and lexical patterns

DM	Precision	Recall	F1-score
Attention	0.31	1	0.47
Bon	0.96	0.91	0.93
La preuve	0.8	1	0.88
Quoi	0.93	0.84	0.88

Table 6 : Results obtained from the ***CORPAIX*** corpus.

6. Evaluation and discussion

Fine-tuning pretrained FlauBERT model for POS tagging task

DM	Precision	Recall	F1-score
Attention	1	0.16	0,27
Bon	1	0.33	0.49

Table 9 : Results of the pre-trained model for DM *attention* and *bon*.

6. Evaluation and discussion

KNN algorithm with FlauBERT word embedding

Corpus	Precision	Recall	F1-score
Attention	0.88	0.88	0.88
Bon	0.99	0.93	0.96
La preuve	-	-	-
Quoi	0.65	0.96	0.78

Table 13 : Results obtained for the categorization of DM in **CORPAIX**.

7. Conclusion

- *DM* cannot be identified only on the basis of the grammatical categories found on its environment.
- The combination of syntactic and lexical information can achieve the goal of identification of the *DM* and *non-DM*.
- Recognizing the DM *attention* seems to be critical to the rule-based approach.
- The machine learning algorithms proves the hypothesis of using word embedding can be beneficial in the case of *attention*.
- Our pre-trained model still affected by scarce and unbalanced data, especially for *attention* and *la preuve*.

7. Perspectives

- Extending the analysis to other polyfunctional DM such as *tiens*, *tu parles*, *bref* and *bon Dieu*.
- Augment the data for the pre-trained model in order to achieve good results and identification.
- Generate an annotated data set in order to help other researchers to contribute in the task.

Thank you !