





# 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

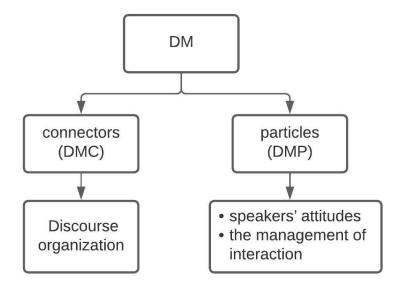
## Plan

- 1. Introduction
- 2. Problem
- 3. Contribution
- 4. Data
- 5. Methodology
- 6. Evaluation and discussion
- 7. Conclusion and Perspectives

# 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

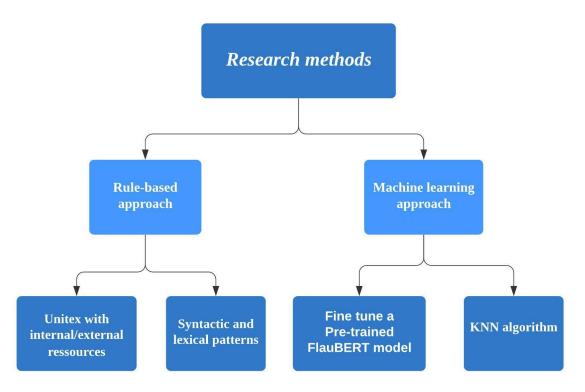
- 1. We built four mechanisms based on rule-based and machine learning.
- 2. We tested different hypothesis and applied several scenarios.
- 3. Made experiment for each mechanism on the same spoken corpora.
- 4. We explored and evaluated the UNITEX platform.

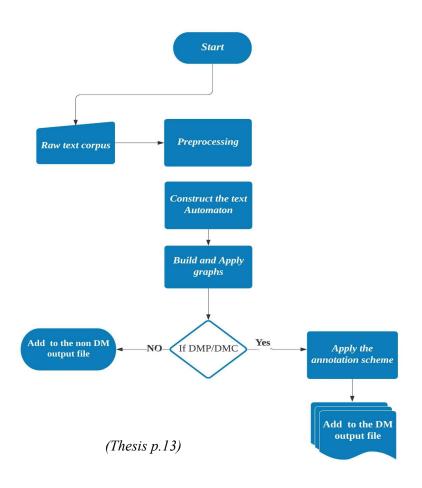
# 4. Data

- 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.

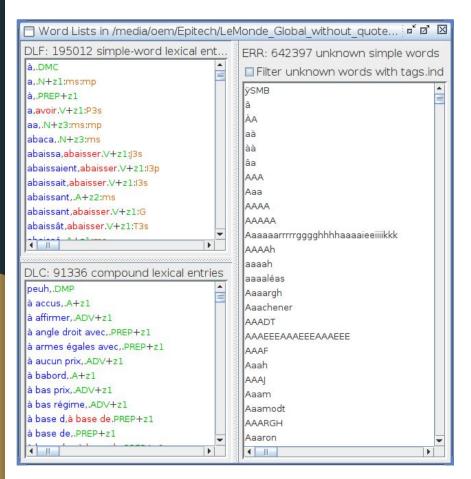
| Corpus  | Attention | bon  | la preuve | quoi |
|---------|-----------|------|-----------|------|
| CORPAIX | 171       | 3867 | 8         | 2380 |
| ESLO    | 163       | 1797 | 11        | 958  |
| TCOF    | 18        | 545  | 0         | 862  |

**Table 1 :** Distribution of DM in the spoken corpus.

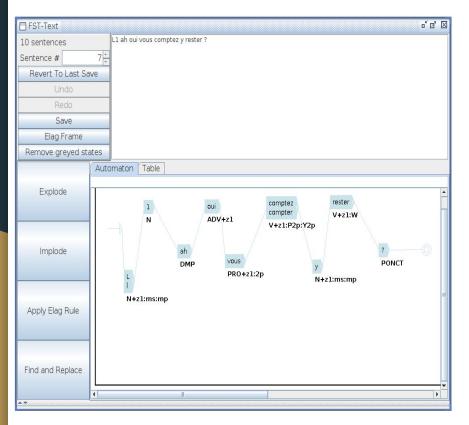




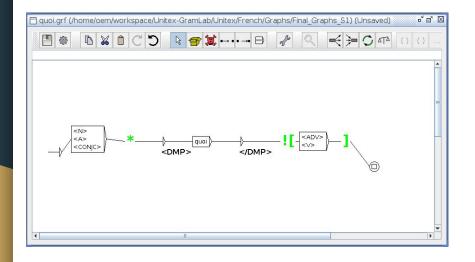
- Read and preprocess the spoken corpus.
- Construct the text automaton.
- Build and apply graphs.
- 4. Visualization of annotation.



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



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



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

### ☐ Concordance: /home/oem/Documents/1.Internship 2021/1.Corpus + dic/Annotation/Ré... 🗝 🗹 🗵 1231 matches

uh pour faciliter l'enfant {S}L2 a <DMP> quoi </DMP> ca servirait sinon {S}L1 mh {S}L3 tu per qu'il {S}L1 ben oui non mais il a <DMP> quoi </DMP> une tête vénusienne martienne quoi donc on parce que le jeudi on a {S}L1 on a <DMP> quoi </DMP> {S}L2 on a partiel sur le roman le jeu e un homme euh on le finit {S}L2 on a <u><DMP> quoi </DMP></u> quarante minutes de pose pour h anglais {S}L1 mh mh mh {S}L2 il y a <DMP> quoi </DMP> {S}L4 relativity {S}L2 ie sais pas eul quand même un apprentissage il y a <u><DMP> quoi </DMP></u> il y a savoir lire et écrire qui est quand vant de dire le mot alors non il y a <DMP> quoi </DMP> il y a si il y a deux trois mots comme ca e mais tu tu as du mal à échapper à ca <DMP> quoi </DMP> hein quand tu travailles avec des gen disons qu'on avait pas trop l'esprit à ca <DMP> quoi </DMP> mais quand j'ai vu qu'il y avait énorn est les les pousser un peu à ça <u><DMP> quoi </DMP></u> moi j'aime bien ce rôle parce que je crois c hts c'est vachement plus sympa ca <DMP> quoi </DMP> {S}L1 tu as de bonnes relations avec t noi ouais je suis un peu maniaque de ça <u><DMP> quoi </DMP></u> {S}L1 et tu es tu penses que tu es a qui ont moi j'ai aimé ça <DMP> quoi </DMP> {S}L ils t'ont touchée {S}L oui ils m'ont touchée on va pas marcher un truc comme ca <DMP> quoi </DMP> ou alors elle va marcher mais move fait on pourrait dire comme ca <DMP> quoi </DMP> on aime bien avoir une belle marque on aime le médecins ou quelque chose comme ca <DMP> quoi </DMP> le corps médical les deux les deu est pas possible de continuer comme ça <DMP> quoi </DMP> on peut pas continuer à faire trav ? mais en fait je suis pas comme ca <DMP> quoi </DMP> mais eh , j'arrive pas j'arrive pas je vais oules Quiès euh des trucs comme ça <DMP> quoi </DMP> il y a rien d'autre {S}L2 mm mais vol eux dire eh ben des trucs comme ca <DMP> quoi </DMP> qui sont c` c` pour moi c` pour moi d vois enfin c'est des trucs comme ça <DMP> quoi </DMP> c'est tu prends les Guignols de l'Info c positions, ou des trucs comme ça <DMP> quoi </DMP> {S}L3 mm mm {S}L1 mais c'est pas l

# 5. Methodology

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

#### Start Construct the text Automaton Raw text corpus **Build and Apply** graphs CamemBERT parser Add to the non DM Yes If DMP/DMC Apply the output file annotation scheme Tagged text corpus Add to the DM output file

(Thesis p.20)

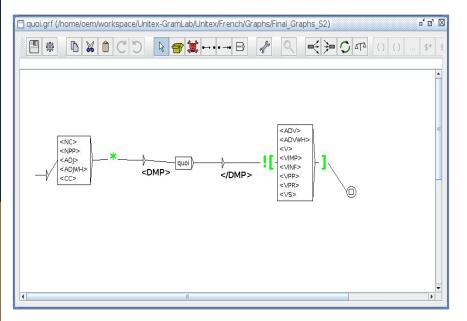
# 5. Methodology

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

#### Corpaix Pos CamemBERT.snt (/home/oem/Documents/1.Internship 2021/1.Corpus + dic/... 💆 🗹 46348 sentence delimiters, 2000006 (38749 diff) tokens, 6150 (15) simple forms, 0 (0) digit {L1,.NC} {et..CC} {pour,.P} {toi,.PRO} {le,.DET} {français,.NC} {c'est..V} {aussi..ADV} {euh,.ADV} {important,.ADJ} {pour,.P} {la,.DET} {communication,.NC} {enfin..PONCT} {c'est..V} {le,.DET} {langage,.NC} {c'est,.V} {important,.AD|} {quoi,.PROWH} {S} {L2,.NC} {pour,.P} {les,.DET}

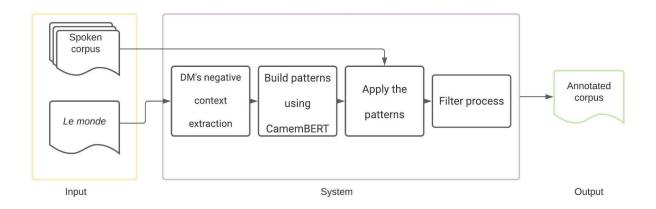
# 5. Methodology

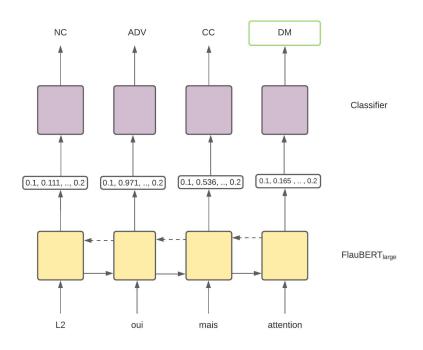
- 1. Tagged text with CamemBERT.
- 2. Unitex reading tagged corpus.
- 3. Build and apply graphs.



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

#### 5.2. Syntactic and lexical patterns

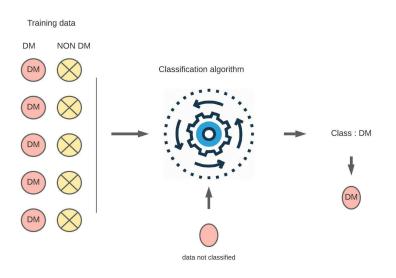




# 5.3. Fine-tuning Pretrained FlauBERT model for PoS Tagging task

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

# 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.

#### 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.

#### 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.

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*.

#### 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!