FOCUS 3
Discours : ressources et systèmes par apprentissage

Jean-Yves Antoine (LIFAT), Agata Savary (LIFAT), Denis Maurel (LIFAT), Anne-Lyse Minard (LLL), Emmanuel Schang (LLL), Lotfi Abouda (LLL), Flora Badin (LLL), Guillaume Cleuziou (LIFO), Sylvie Billot (LIFO), Gaëtan Caillaut (LIFO), Anaïs Lefeuvre- Halftermeyer (LIFO), Loïc Grobol (LATTICE)

Plan

Phénomènes

Données

Méthodes

Outils

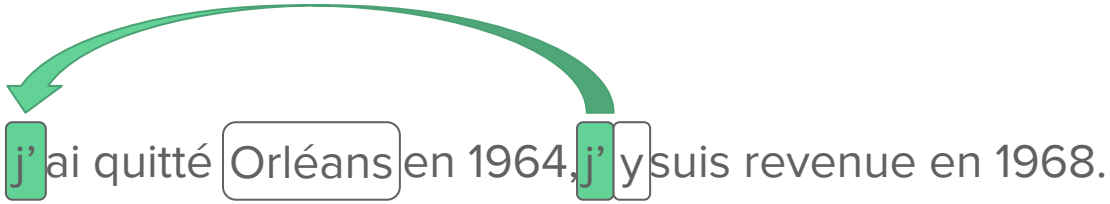
Phénomènes

- Coréférences
- Temporalité

j’ai quitté Orléans en 1964, j’y suis revenue en 1968.

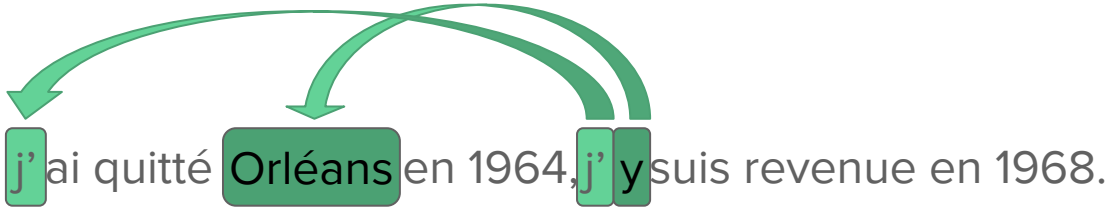
Phénomènes

- Coréférences
- Temporalité



Phénomènes

- **Coréférences**
- Temporalité

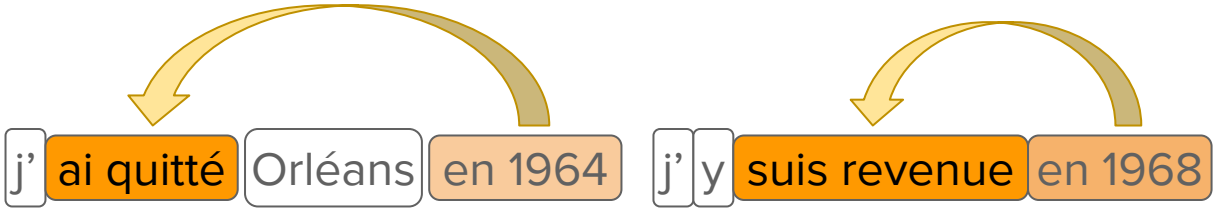


Phénomènes

- Coréférences
- Temporalité

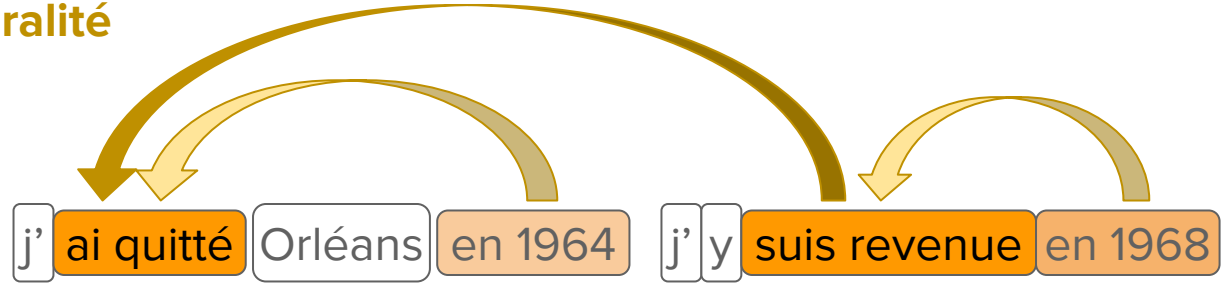
Phénomènes

- Coréférences
- Temporalité



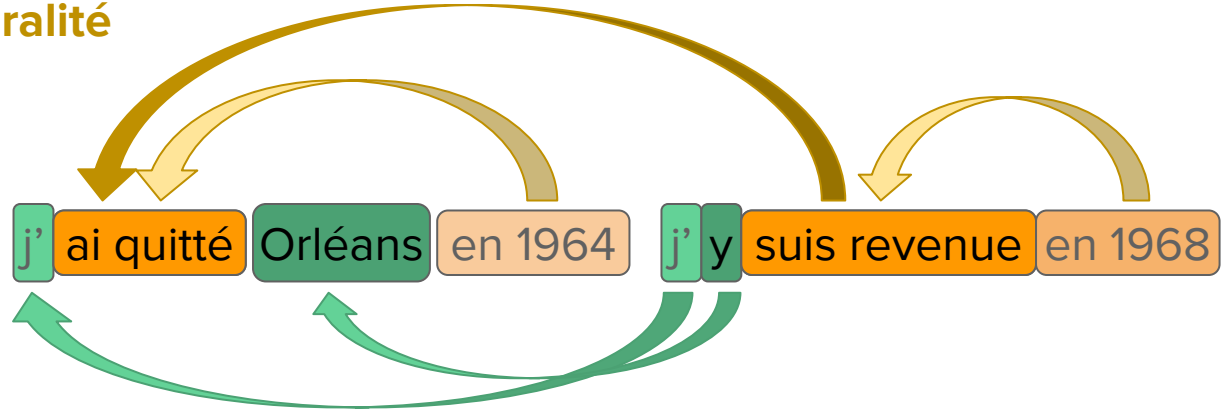
Phénomènes

- Coréférences
- **Temporalité**



Phénomènes

- **Coréférences**
- **Temporalité**



Données

- Corpus ANCOR

  oral transcrit, spontané, avec interactivité variable

  ⇒ plus grand corpus de la coréférence à l'oral, plus gros corpus français annoté en coréférence = 488 000 mots (jusqu'à DEMOCRAT, imminent)

- Corpus ODIL

  en cours de réalisation, autre couche du corpus ANCOR, annoté en syntaxe automatiquement, temporalité (semi ?) manuellement

Données

- Corpus ANCOR

  Annotation manuelle de traits fins :

  - genre,
  - nombre,
  - type d'EN,
  - définitude,
  - généricité,
  - etc.

- Corpus ODIL

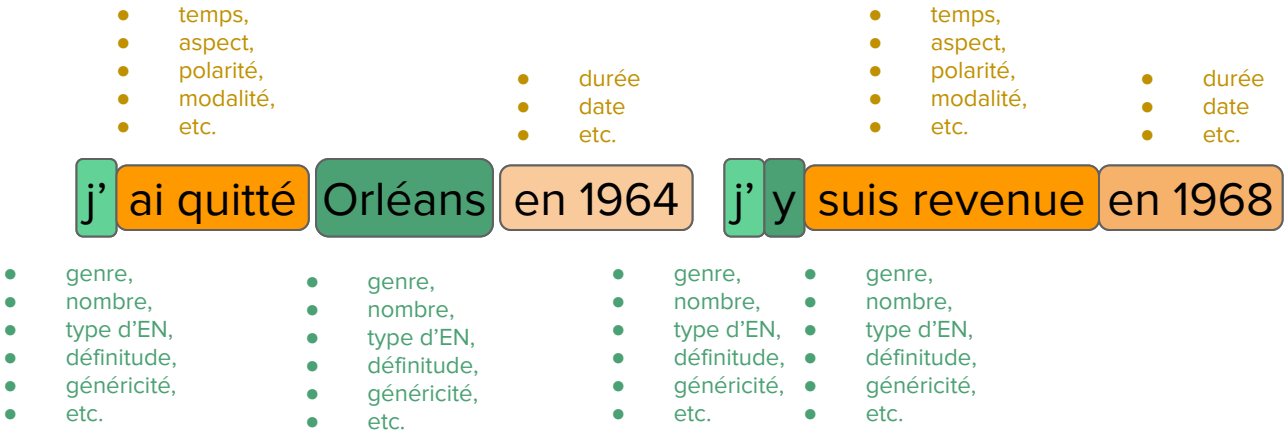
  Syntaxe automatique

  Correction manuelle et annotation manuelle de traits fins :

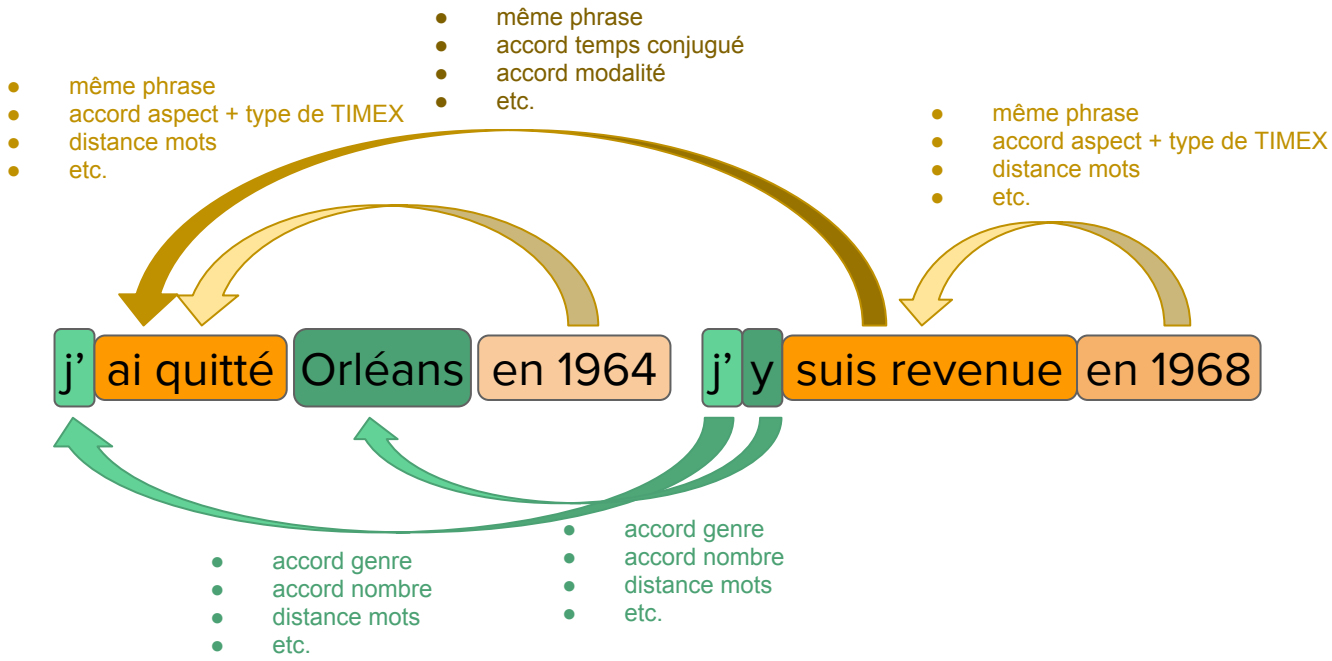
  - temps,
  - aspect,
  - polarité,
  - modalité,
  - etc.

Données

Données


temps,
aspect,
polarité,
modalité,
etc.
durée
date
etc.
temps,
aspect,
polarité,
modalité,
etc.
durée
date
etc.
j’
ai quitté
Orléans
en 1964
j’
y
suis revenue
en 1968
genre,
nombre,
type d’EN,
définitude,
généricité,
etc.
genre,
nombre,
type d’EN,
définitude,
généricité,
etc.
genre,
nombre,
type d’EN,
définitude,
généricité,
etc.
genre,
nombre,
type d’EN,
définitude,
généricité,
etc.

Données


même phrase
accord temps conjugué
accord modalité
etc.
même phrase
accord aspect + type de TIMEX
distance mots
etc.
même phrase
accord aspect + type de TIMEX
distance mots
etc.
j’
ai quitté
Orléans
en 1964
j’
y
suis revenue
en 1968
accord genre
accord nombre
distance mots
etc.
accord genre
accord nombre
distance mots
etc.

Méthodes

- Travail sur la norme ISO-TIMEML
- Production de corpus annotés

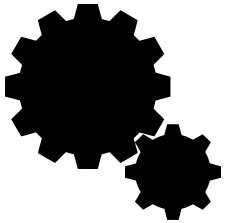
- Apprentissage supervisé pour la création de chaînes référentielles : Naive Bayes, Decision Tree, SVM

- Apprentissage d’espace prétopologique structurant pour extraire un graphe temporel issu d’un discours (Gaëtan Caillaut LIFO)

Méthodes

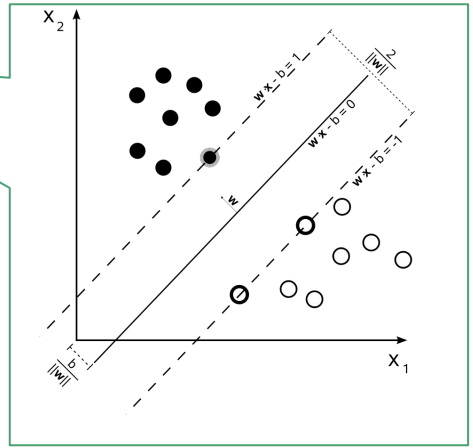
corpus ANCOR

modèle

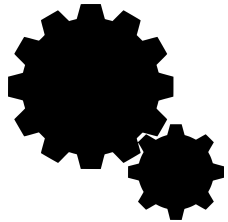


- **Apprentissage supervisé** pour la création de chaînes référentielles : Naive Bayes, **Decision Tree**, **SVM**

corpus ANCOR



modèle



Méthodes

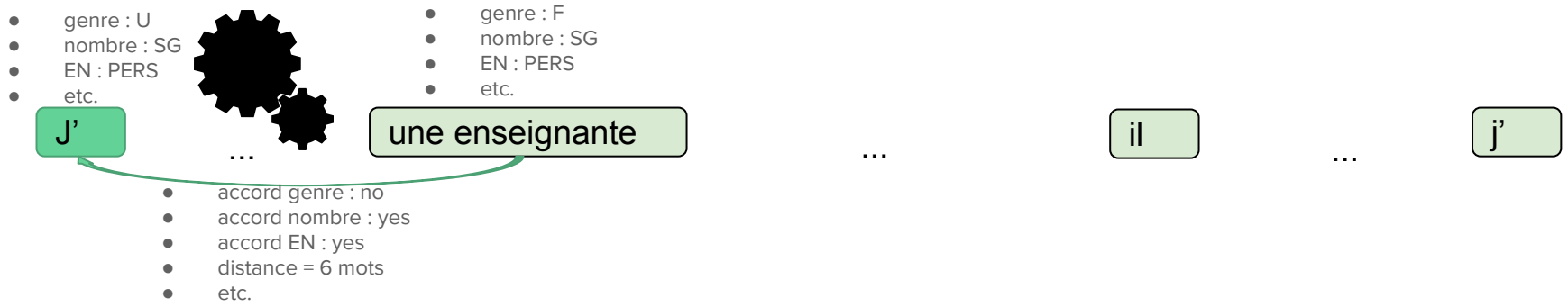
- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

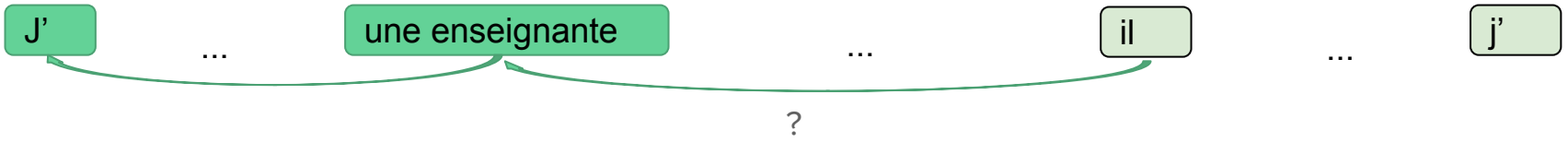


Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

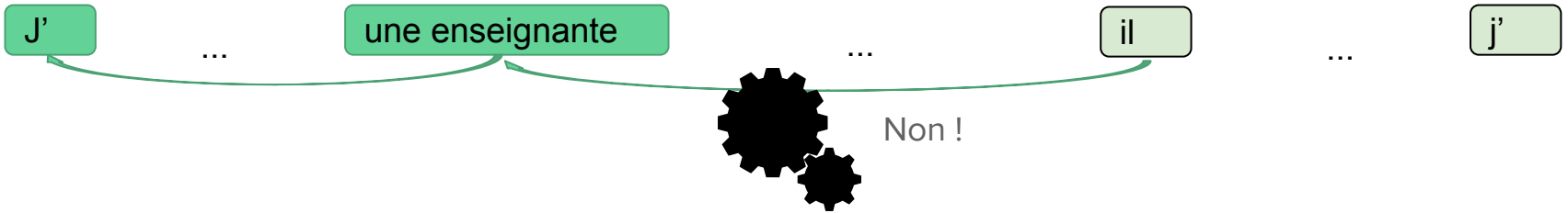
Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM



Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

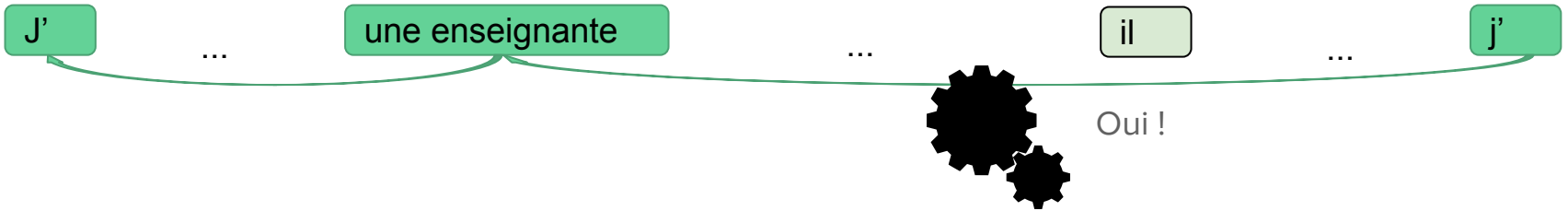


Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM

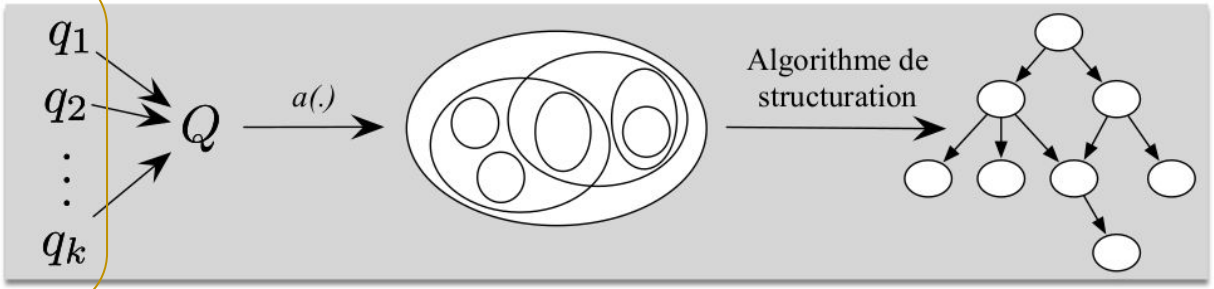
Méthodes

- Apprentissage supervisé pour la **création de chaînes référentielles** : Naive Bayes, Decision Tree, SVM



Méthodes

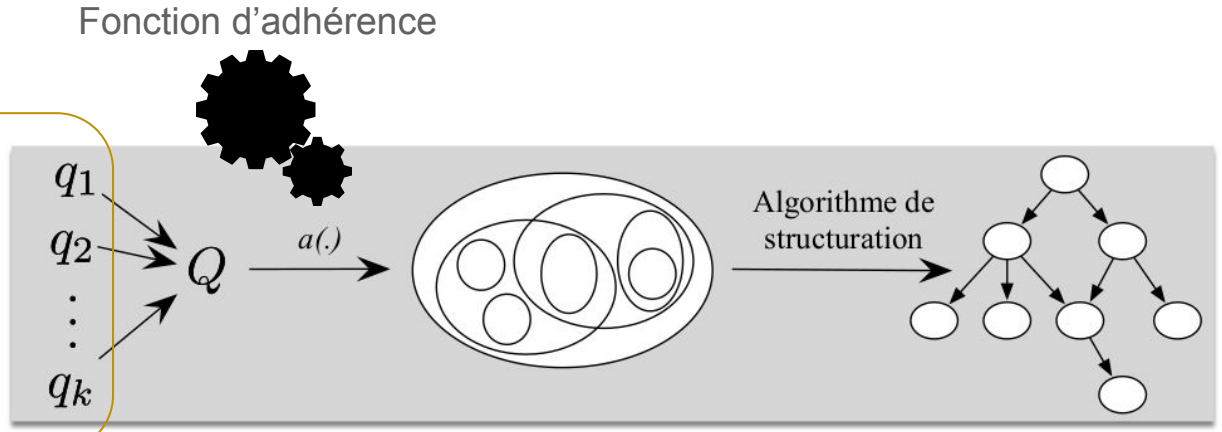
Formule logique de combinaison de “features”



- Apprentissage d’espace prétopologique structurant pour extraire un graphe temporel issu d’un discours (Gaëtan Caillaut LIFO)

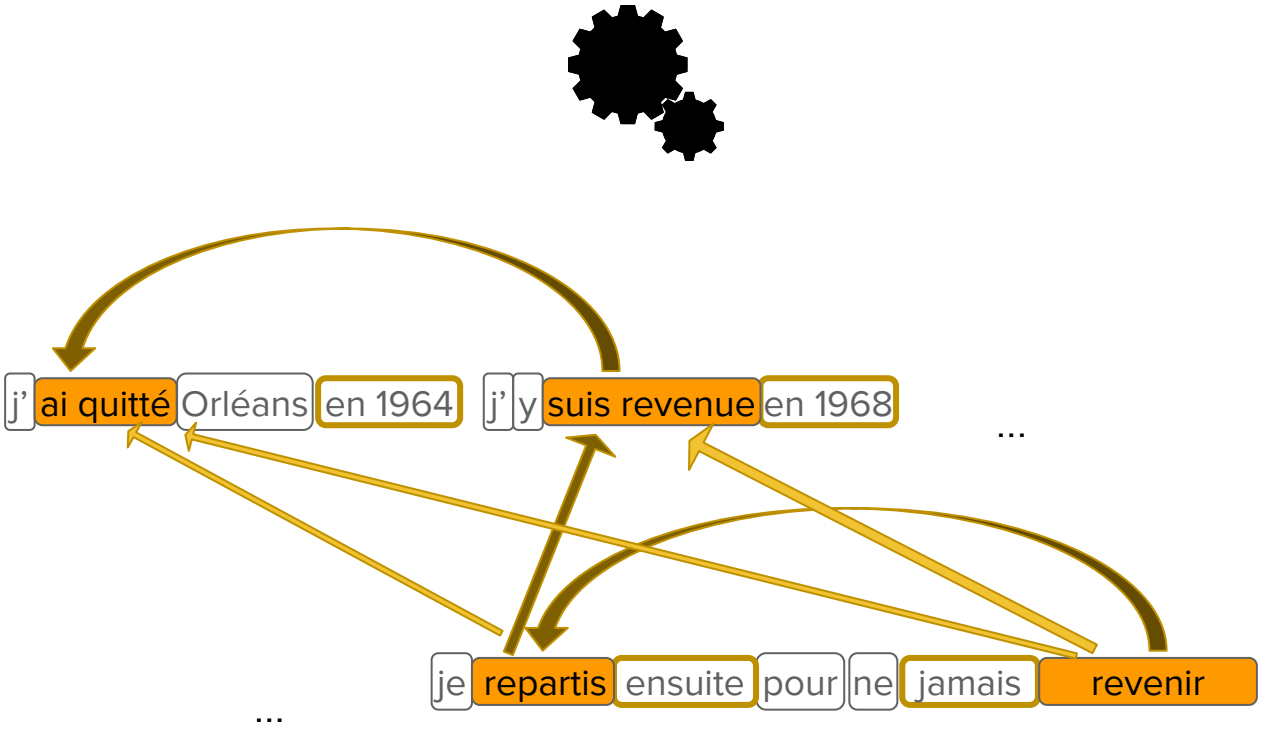
Méthodes

Formule logique de combinaison de “features”

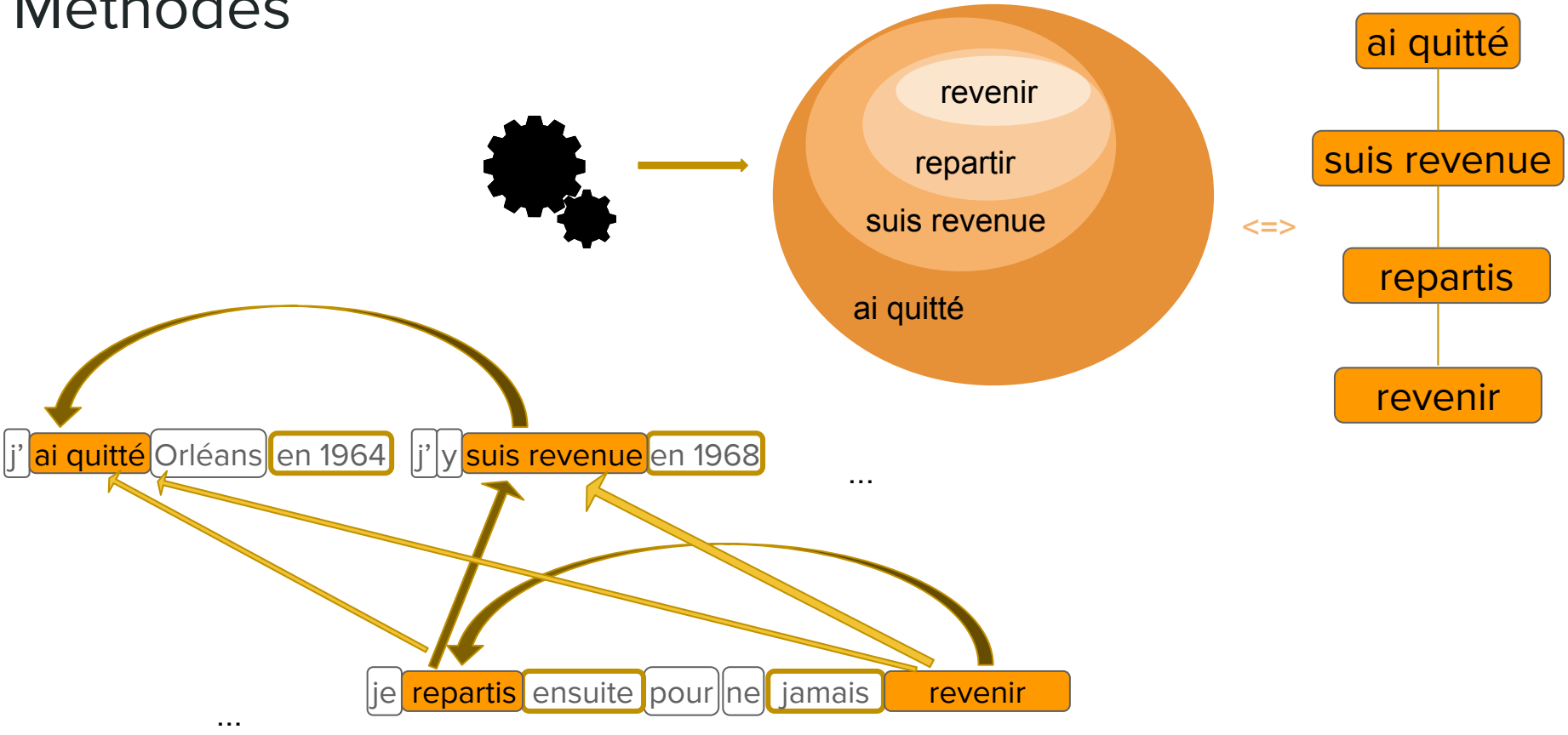


- Apprentissage d’espace prétopologique structurant pour extraire un graphe temporel issu d’un discours (Gaëtan Caillaut LIFO)

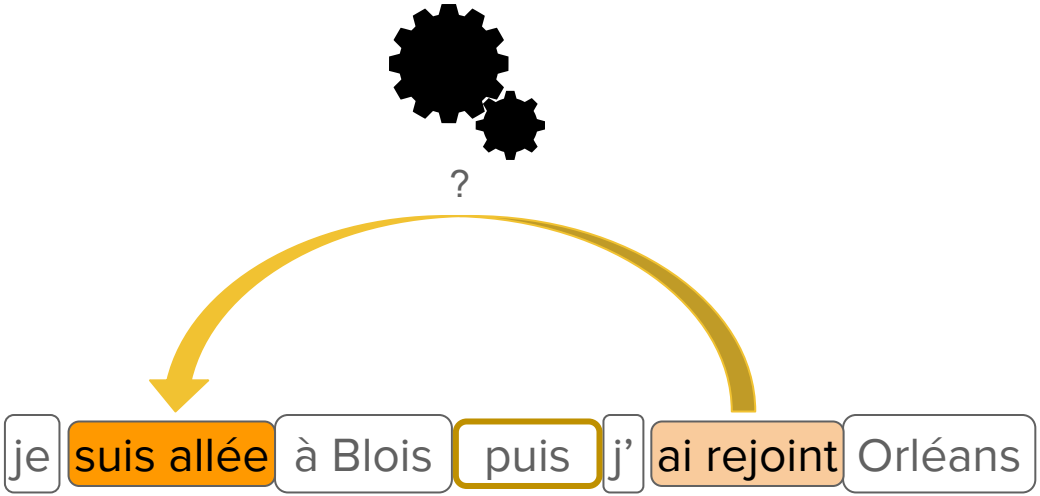
Méthodes


j’ ai quitté Orléans en 1964
j’ y suis revenue en 1968
...
je repartis ensuite pour ne jamais revenir
...

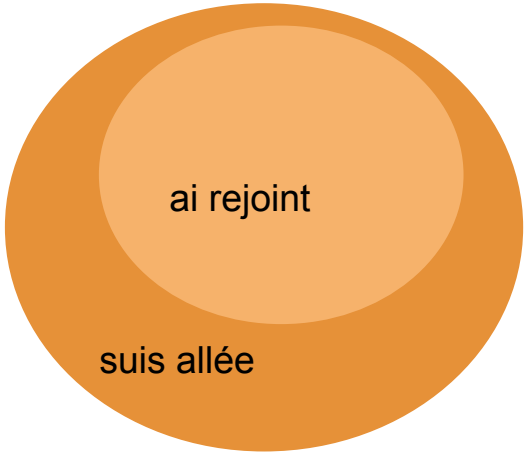
Méthodes


revenir
repartir
suis revenue
ai quitté
<=>
ai quitté
suis revenue
repartis
revenir
j’ ai quitté Orléans en 1964
j’ y suis revenue en 1968
...
je repartis ensuite pour ne jamais revenir
...

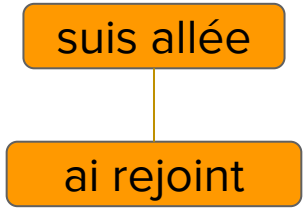
Méthodes

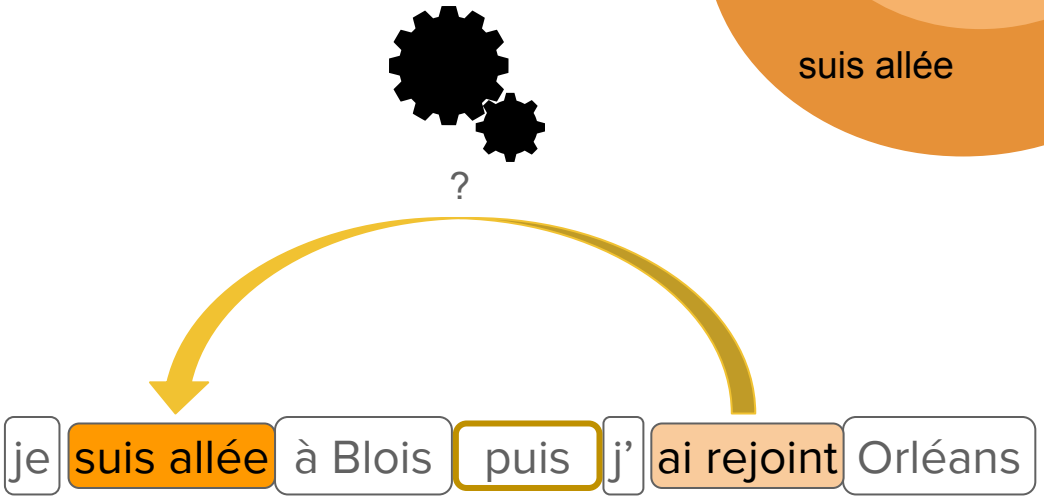


Méthodes

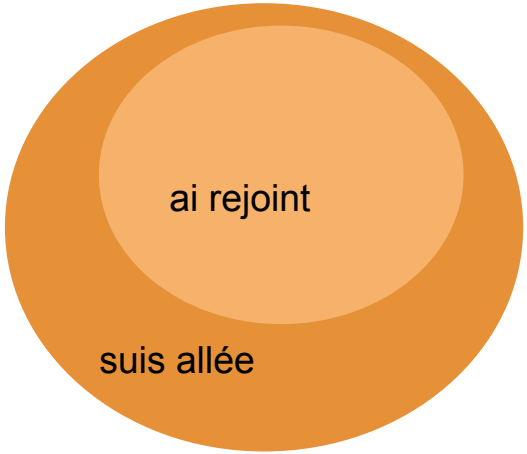

ai rejoint
suis allée

<=>

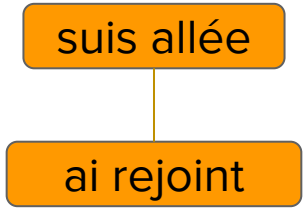

suis allée
ai rejoint

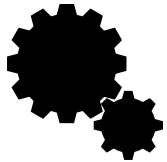

?
je
suis allée
à Blois
puis
j’
ai rejoint
Orléans

Méthodes

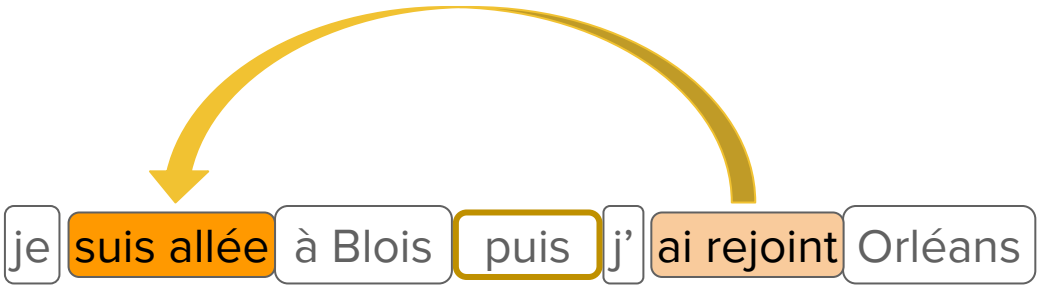

ai rejoint
suis allée

<=>


suis allée
ai rejoint



<


je
suis allée
à Blois
puis
j'
ai rejoint
Orléans

Outils

- **RÉSOLUTION :** CROC (LATTICE)... et ses petits ! (collaboration Loïc Grobol LATTICE)
  - rendre la chaine end-to-end
  - intégration de la détection de mention

- **RÉSOLUTION :** PrétopoTal
  - adaptation de la création d'ontologies à la création de graphes temporels de discours

- **ANNOTATION :** CONTEMPLATA
  - refonte annotation

Autres collaborations

RAVIOLI

Métriques

PREDICT4ALL

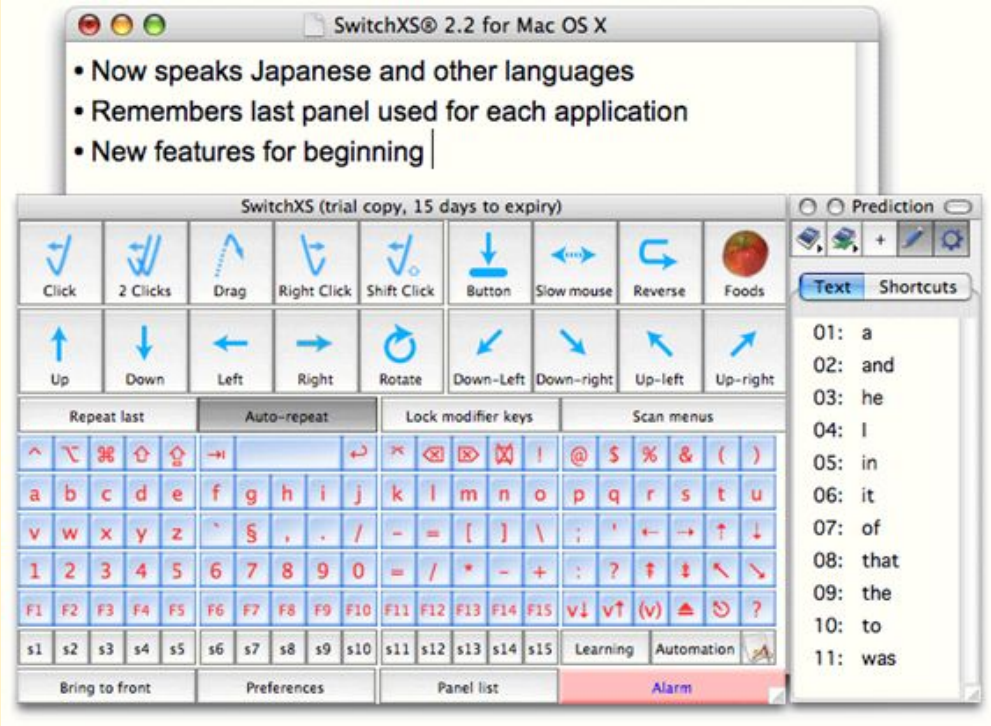
PREDICT4ALL (aide au handicap)

Phénomènes et Données

Méthodes et Outils

Phénomènes et données

- Saisie de texte pour personnes souffrant de handicaps moteurs sévères (locked-in syndrom, paralysie cérébrale, maladie de Charcot, tétraplégie...)
- **Clavier virtuel**
- **Enjeu TAL** : prédiction de texte pour saisie plus rapide



Méthodes

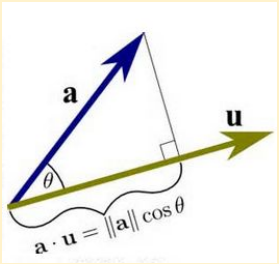
- Prédiction de texte combinant un modèle de langage statistique (syntaxe) et un modèle d'adaptation sémantique (thématique du discours)

$$P'(w_i) = \frac{P_b(w_i)^{\lambda_1} \cdot P_s(w_i)^{(1-\lambda_1)}}{\sum_{j=1}^{n} P_b(w_j)^{\lambda_1} \cdot P_s(w_j)^{(1-\lambda_1)}}$$

$$P(w_i) = \lambda_1 \times P_{Base}(w_i \mid w_{i-1}\, w_{i-2}\, w_{i-3}) + (1-\lambda_1) \times P_{DUM}(w_i \mid w_{i-1}\, w_{i-2})$$

N-gram général +
N-gram utilisateur

$$P_{LSA}(w_i|h) = \frac{\left(\cos(\vec{w}_i, \vec{h}) - \cos_{min}(\vec{h})\right)^{\gamma}}{\sum_k \left(\cos(\vec{w}_k, \vec{h}) - \cos_{min}(\vec{h})\right)^{\gamma}}$$



Analyse sémantique latente
(équivalence *word embeddings*)

=> espace vectoriel sémantique

Outils

- Système Sibylle vK : diffusion libre (*open source* en 2020)

- **Predict4All** adaptation troubles dyslexiques