Exploration de règles d'annotation pour la reconnaissance d'entités nommées

Damien Nouvel

Jean-Yves Antoine(directeur)

Nathalie Friburger, Arnaud Soulet (encadrants)

Université François Rabelais Tours
Laboratoire d'Informatique
Equipe BDTLN

Entités nommées

Exemples d'entités nommées

*'Le **20 mai 2010**, pour **13 euros**, il avait visité le **Centre Georges Pompidou** à **Paris**. Il avait beaucoup aimé les œuvres de l'**artiste-peintre Salvador Dali** et s'était promis de voir le film **Un chien andalou** dès que possible.'*

Entités nommées

- Extraction d'informations au sein du langage naturel

Exemples d'entités nommées

*'Le **20 mai 2010**, pour **13 euros**, il avait visité le **Centre Georges Pompidou** à **Paris**. Il avait beaucoup aimé les œuvres de l'**artiste-peintre Salvador Dali** et s'était promis de voir le film **Un chien andalou** dès que possible.'*

Entités nommées

- Extraction d'informations au sein du langage naturel
- Deux types d'expressions linguistiques concernées :
  - **Noms propres** : personnes, lieux, organisations, produits
  - **Descriptions définies** : expressions de temps, montants, fonctions

Exemples d'entités nommées

*'Le **20 mai 2010**, pour **13 euros**, il avait visité le **Centre Georges Pompidou** à **Paris**. Il avait beaucoup aimé les œuvres de l'**artiste-peintre Salvador Dali** et s'était promis de voir le film **Un chien andalou** dès que possible.'*

Entités nommées

- Extraction d'informations au sein du langage naturel
- Deux types d'expressions linguistiques concernées :
  - **Noms propres** : personnes, lieux, organisations, produits
  - **Descriptions définies** : expressions de temps, montants, fonctions

Exemples d'entités nommées

*'Le **20 mai 2010**, pour **13 euros**, il avait visité le **Centre Georges Pompidou** à **Paris**. Il avait beaucoup aimé les œuvres de l'**artiste-peintre Salvador Dali** et s'était promis de voir le film **Un chien andalou** dès que possible.'*

$\Rightarrow$ *Comment différencier les entités nommées dans les textes ?*

Contexte applicatif

Contexte applicatif

- Utilisation des entités nommées :
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - ...

Contexte applicatif

- Utilisation des entités nommées :
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - ...

- Collecter des informations sur les entités nommées :

Contexte applicatif

- Utilisation des entités nommées :
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - ...

- Collecter des informations sur les entités nommées :
  - **Détection** : indiquer où sont les entités nommées

Contexte applicatif

- Utilisation des entités nommées :
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - ...

- Collecter des informations sur les entités nommées :
  - **Détection** : indiquer où sont les entités nommées
  - **Reconnaissance** : préciser leur type (personne, ville, société, etc.)

Contexte applicatif

- Utilisation des entités nommées :
  - **Indexation et recherche d'informations**
  - **Question - réponse**
  - **Annotation en rôles sémantiques**
  - **Résolution d'autres tâches** (transcription, syntaxe, anaphores)
  - ...

- Collecter des informations sur les entités nommées :
  - **Détection** : indiquer où sont les entités nommées
  - **Reconnaissance** : préciser leur type (personne, ville, société, etc.)
  - **Résolution** : déterminer toutes leurs propriétés (référence, valeur)

Difficultés

- Phénomènes liés à la langue naturelle :

Difficultés

- Phénomènes liés à la langue naturelle :
  - **Synonymie** : référent désigné par de nombreuses expressions

Exemples ambigus

*'Parmi les pères fondateurs, **Washington** est le plus connu.'* (*Personne*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*Personne*)

Difficultés

- Phénomènes liés à la langue naturelle :
  - **Synonymie** : référent désigné par de nombreuses expressions
  - **Homonymie** : même expression désignant plusieurs référents

Exemples ambigus

*'Parmi les pères fondateurs, **Washington** est le plus connu.'* (*Personne*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*Personne*)
*'Nous avons été en vacances à **Washington** et à Boston.'* (*Ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*Paquebot*)

Difficultés

- Phénomènes liés à la langue naturelle :
  - **Synonymie** : référent désigné par de nombreuses expressions
  - **Homonymie** : même expression désignant plusieurs référents
  - **Métonymie** : interprétation sémantique d'une expression

Exemples ambigus

*'Parmi les pères fondateurs, **Washington** est le plus connu.'* (*Personne*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*Personne*)
*'Nous avons été en vacances à **Washington** et à Boston.'* (*Ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*Paquebot*)
*'Hier, **Washington** a battu New York 34 à 10.'* (*Equipe sportive*)
*'Lors des discussions, **Washington** a opposé son veto.'* (*Gouv.*)

Difficultés

- Phénomènes liés à la langue naturelle :
  - **Synonymie** : référent désigné par de nombreuses expressions
  - **Homonymie** : même expression désignant plusieurs référents
  - **Métonymie** : interprétation sémantique d'une expression

Exemples ambigus

*'Parmi les pères fondateurs, **Washington** est le plus connu.'* (*Personne*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*Personne*)
*'Nous avons été en vacances à **Washington** et à Boston.'* (*Ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*Paquebot*)
*'Hier, **Washington** a battu New York 34 à 10.'* (*Equipe sportive*)
*'Lors des discussions, **Washington** a opposé son veto.'* (*Gouv.*)

$\Rightarrow$ *Comment traiter automatiquement ce cas ?*

Plan

Définition des entités nommées ?

- A propos des entités nommées :

Définition des entités nommées ?

- A propos des entités nommées :
  - *“Identifying the names of all the people, organizations and geographic locations in a text”* [Grishman & Sundheim 1996]

Définition des entités nommées ?

- A propos des entités nommées :
  - *“Identifying the names of all the people, organizations and geographic locations in a text”* [Grishman & Sundheim 1996]
  - *“Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.”* [Ehrmann 2008]

Définition des entités nommées ?

- A propos des entités nommées :
  - *"Identifying the names of all the people, organizations and geographic locations in a text"* [Grishman & Sundheim 1996]
  - *"Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus."* [Ehrmann 2008]
  - *"Mono- or multi-word expression belonging to a potentially interesting class for an application"* [Galibert et. al. 2011]

Définition des entités nommées ?

- A propos des entités nommées :
  - *"Identifying the names of all the people, organizations and geographic locations in a text"* [Grishman & Sundheim 1996]
  - *"Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus."* [Ehrmann 2008]
  - *"Mono- or multi-word expression belonging to a potentially interesting class for an application"* [Galibert et. al. 2011]
  - *"Les entités nommées incluent traditionnellement trois grandes classes : les noms, les quantités, les dates et durées. Nous nous plaçons dans le contexte d'extraction d'informations (entités, relations) servant à constituer une base de connaissances."* [Rosset et. al. 2011]

Définition des entités nommées ?

- A propos des entités nommées :
  - *"Identifying the names of all the people, organizations and geographic locations in a text"* [Grishman & Sundheim 1996]
  - *"Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus."* [Ehrmann 2008]
  - *"Mono- or multi-word expression belonging to a potentially interesting class for an application"* [Galibert et. al. 2011]
  - *"Les entités nommées incluent traditionnellement trois grandes classes : les noms, les quantités, les dates et durées. Nous nous plaçons dans le contexte d'extraction d'informations (entités, relations) servant à constituer une base de connaissances."* [Rosset et. al. 2011]

$\Rightarrow$ *Définitions applicatives ou en extension*

Les campagnes d'évaluation

- Compétitions en recherche d'information

Campagne	Langue	Modalité	Types
MUC-6 (96)	Anglais	Ecrit (rapports)	pers, org, loc
MUC-7 (97)	Anglais	Ecrit	MUC-6 + date, heure, montant, pourcent
MET-1 (97)	Espagnol, chinois, japonais	Ecrit	MUC-7
IREX (99)	Japonais	Ecrit	MUC-7 + artefact
CoNLL (02)	Espagnol, flamand	Ecrit	MUC-6 + misc
ACE (07)	Anglais, arabe, chinois	Ecrit (dont spont.)	MUC-6 + bâtiments, entité géo-politique, armes, véhicules
ESTER2 (09)	Français	Oral	MUC-6 + temps, montant, fonction, produit
EVALITA (11)	Italien	Oral	MUC-6 + entité géo-politique
ETAPE (12)	Français	Oral (dont spont.)	ESTER2 + évènements

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

- Annotations Quaero, avec imbrications (composants) :

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

- Annotations Quaero, avec imbrications (composants) :
  - **Entités** : personne (`pers`), fonction (`fonc`), organisation (`org`), lieu (`loc`), production humaine (`prod`), point dans le temps (`time`), quantités (`amount`), évènements (`event`)

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

- Annotations Quaero, avec imbrications (composants) :
  - **Entités** : personne (`pers`), fonction (`fonc`), organisation (`org`), lieu (`loc`), production humaine (`prod`), point dans le temps (`time`), quantités (`amount`), évènements (`event`)
  - **Composants** : genre, objet, valeur (transversales) ; prénom, nom (`pers`) ; jour, mois, année (`time`), unité monétaire (`amount`), etc.

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

- Annotations Quaero, avec imbrications (composants) :
  - **Entités** : personne (`pers`), fonction (`fonc`), organisation (`org`), lieu (`loc`), production humaine (`prod`), point dans le temps (`time`), quantités (`amount`), évènements (`event`)
  - **Composants** : genre, objet, valeur (transversales) ; prénom, nom (`pers`) ; jour, mois, année (`time`), unité monétaire (`amount`), etc.

$\Rightarrow$ *Transcriptions manuelles et automatiques (WER>20)*

Campagne ETAPE

- Campagne d'évaluation faisant suite à ESTER2 (AFCP) :
  - **Segmentation** : locuteurs, jingles, musique, bruits de fond, etc.
  - **Transcription** : reconnaissance de la parole (signal) en texte
  - **Recherche d'informations** : entités nommées

- Annotations Quaero, avec imbrications (composants) :
  - **Entités** : personne (`pers`), fonction (`fonc`), organisation (`org`), lieu (`loc`), production humaine (`prod`), point dans le temps (`time`), quantités (`amount`), évènements (`event`)
  - **Composants** : genre, objet, valeur (transversales) ; prénom, nom (`pers`) ; jour, mois, année (`time`), unité monétaire (`amount`), etc.

$\Rightarrow$ *Transcriptions manuelles et automatiques (WER>20)*

$\Rightarrow$ *Annotation manuelle, projection et évaluation*

Corpus ETAPE

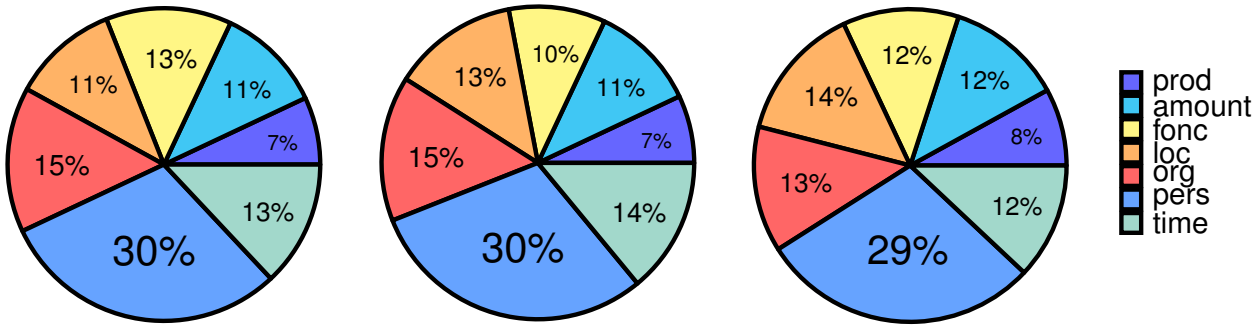
- 74 enregistrements : BFMTV, France Inter, LCP, TV8 (hors Quaero)

Corpus	Tokens	Enoncés	EN
Etape-Train	355 975	14 989	46 259
Etape-Dev	115 530	5 724	14 112
Etape-Test	123 221	6 770	13 055
Total	594 726	27 483	73 426

Corpus ETAPE

- 74 enregistrements : BFMTV, France Inter, LCP, TV8 (hors Quaero)

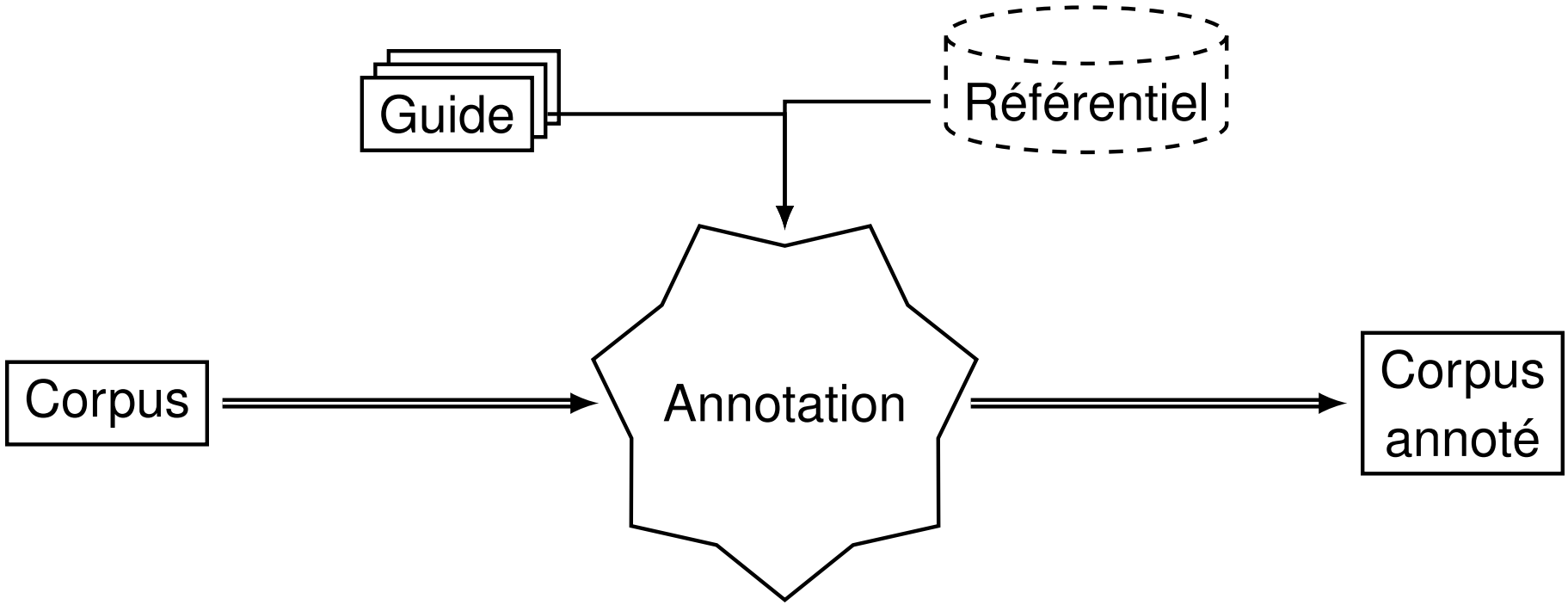
Corpus	Tokens	Enoncés	EN
Etape-Train	355 975	14 989	46 259
Etape-Dev	115 530	5 724	14 112
Etape-Test	123 221	6 770	13 055
Total	594 726	27 483	73 426



(d) Etape-Train (e) Etape-Dev (f) Etape-Test

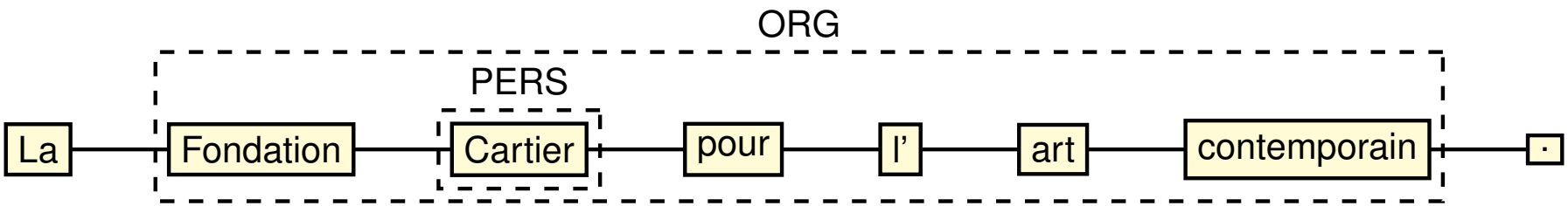
Annotation de corpus

Annotation de corpus



- **Corpus** : ensemble de textes, fichiers à annoter
- **Guide** : directives concernant les types et structures à annoter

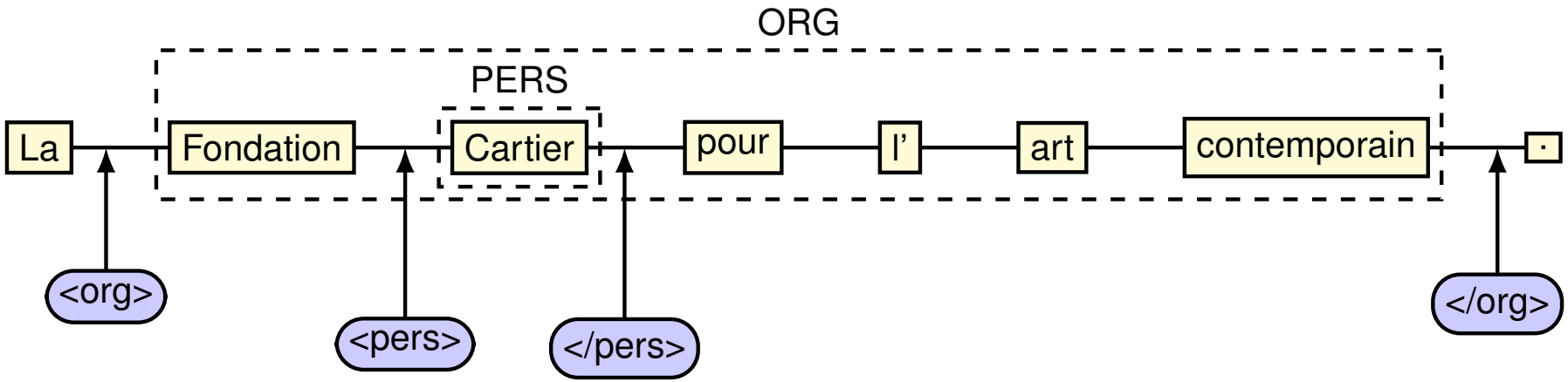
Représentations des annotations



Représentations des annotations

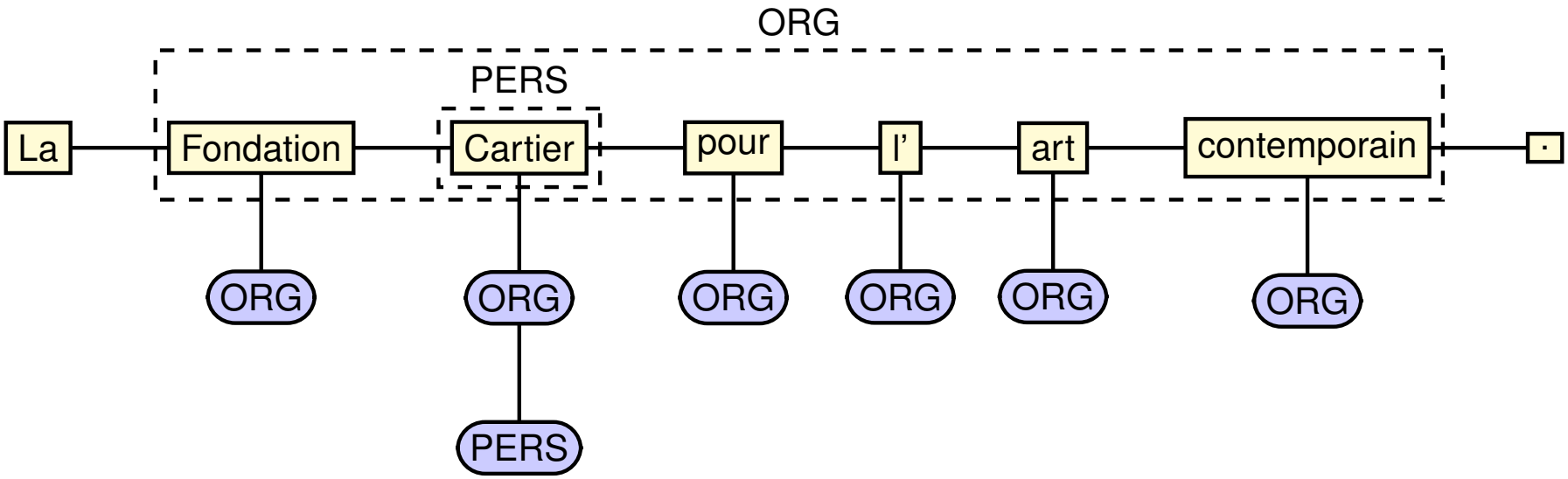
- **Balises d'annotation (marqueurs)** :
  - `'<org> fondation <pers> Cartier </pers> </org>'`

    ⇒ *Balises **séparées** en début et fin d'annotation*



Représentations des annotations

- **Balises d'annotation (marqueurs)** :
  - `'<org> fondation <pers> Cartier </pers> </org>'`
    ⇒ *Balises **séparées** en début et fin d'annotation*
- **Classes de mots** :
  - `'fondation/ORG Cartier/PERS+ORG'`
    ⇒ *Attribue aux **données (mots)** des **classe***



Les marqueurs d'annotation

Annotation en composants

‘ Le <date> 20 mai 2010 </date>, pour <montant> 13 euros </montant>, il avait visité le <org> centre Georges Pompidou </org> à <loc> Paris </loc>. ’

Les marqueurs d'annotation

Annotation en composants

‘ Le <date> <jour> 20 </jour> <mois> mai </mois> <annee> 2010 </annee> </date>, pour <montant> <num> 13 </num> <monnaie> euros </monnaie> </montant>, il avait visité le <org> centre <pers> <prenom> Georges </prenom> <nom> Pompidou </nom> </pers> </org> à <loc> Paris </loc>. ’

Les marqueurs d'annotation

Annotation en composants

```
‘ Le <date> <jour> 20 </jour> <mois> mai </mois> <annee> 2010
</annee> </date>, pour <montant> <num> 13 </num> <monnaie>
euros </monnaie> </montant>, il avait visité le <org> centre
<pers> <prenom> Georges </prenom> <nom> Pompidou </nom>
</pers> </org> à <loc> Paris </loc>. ’
```

- **Marqueurs** comme instructions de structuration locale
  - Prise en compte du **contexte local** (séquences)
  - Besoin **limité** d'inférences complexes

  $\Rightarrow$ *Modèle pour ajouter un marqueur **individuel** ?*

Les marqueurs d'annotation

Annotation en composants

```
‘ Le <date> <jour> 20 </jour> <mois> mai </mois> <annee> 2010
</annee> </date>, pour <montant> <num> 13 </num> <monnaie>
euros </monnaie> </montant>, il avait visité le <org> centre
<pers> <prenom> Georges </prenom> <nom> Pompidou </nom>
</pers> </org> à <loc> Paris </loc>. ’
```

- **Marqueurs** comme instructions de structuration locale
  - Prise en compte du **contexte local** (séquences)
  - Besoin **limité** d'inférences complexes

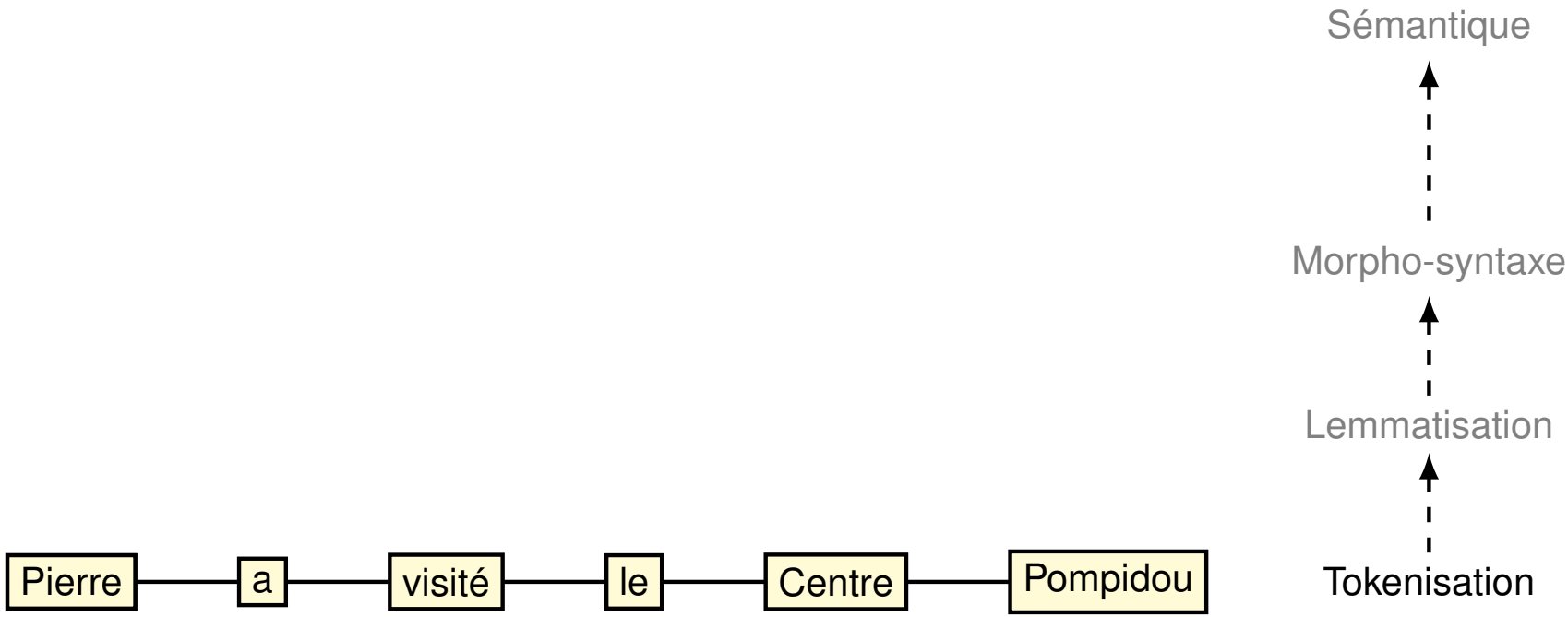
  $\Rightarrow$ *Modèle pour ajouter un marqueur **individuel** ?*
- Rétrospective des théories linguistiques

$\Rightarrow$ *Proposition de propriétés de **stabilité** et d'**opérabilité***

Plan

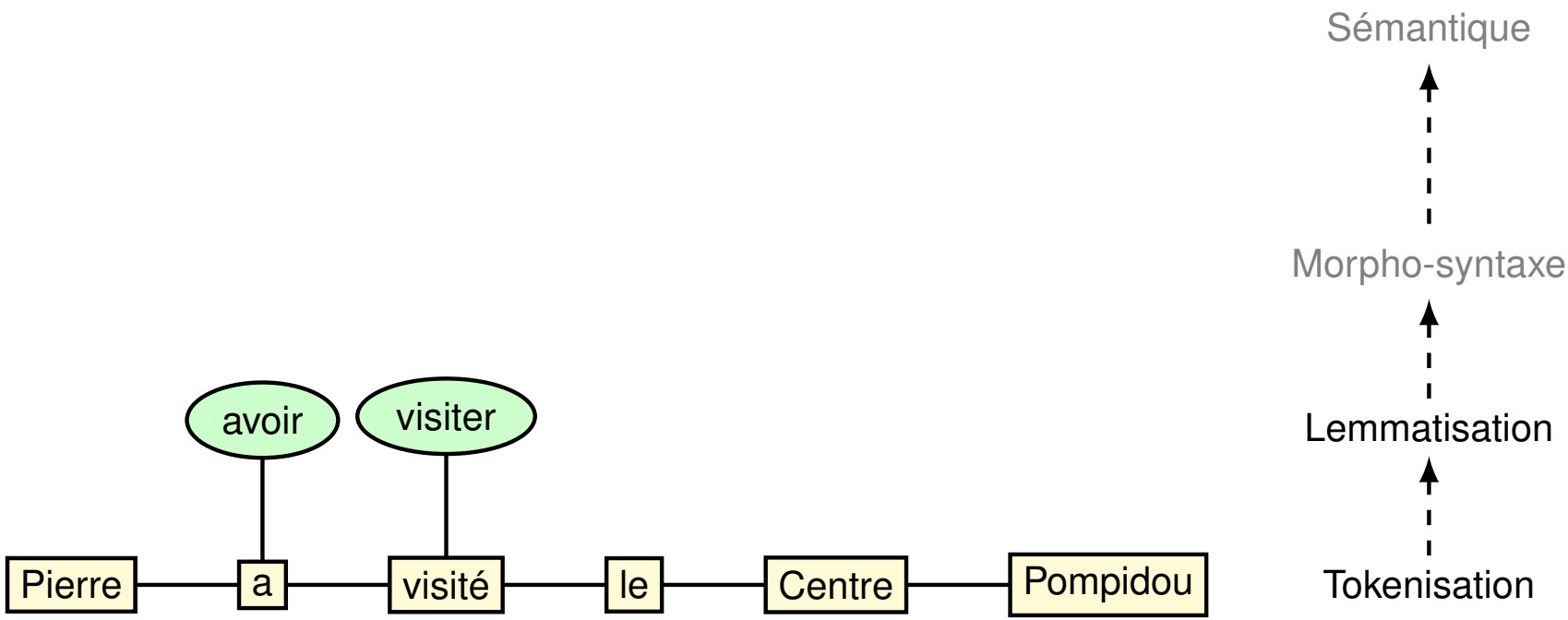
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse :



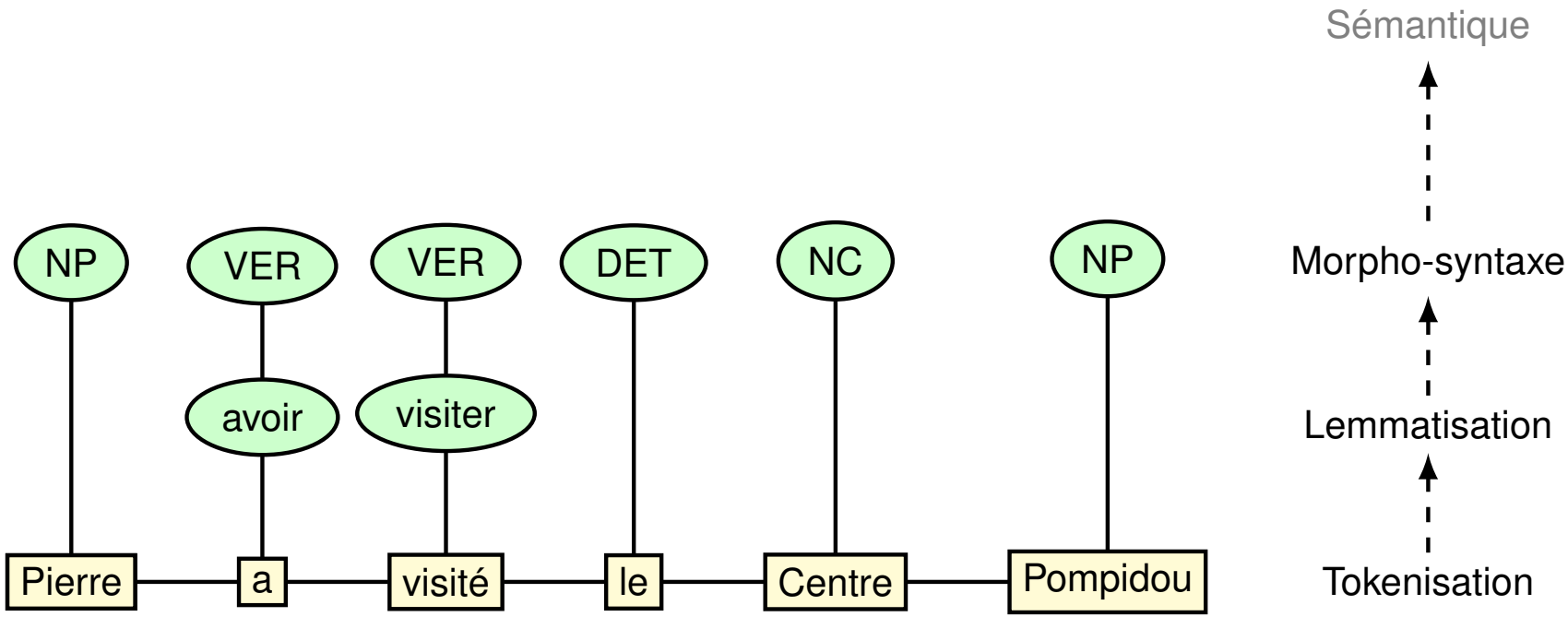
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse :



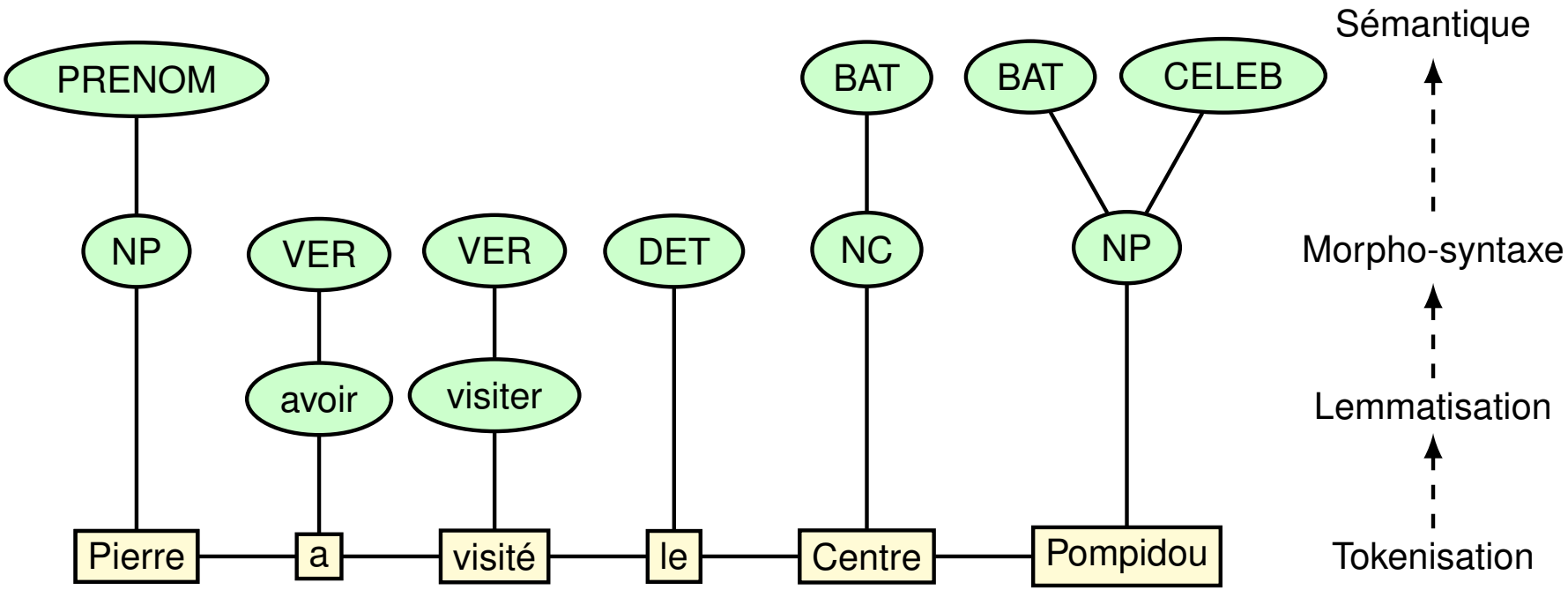
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse :



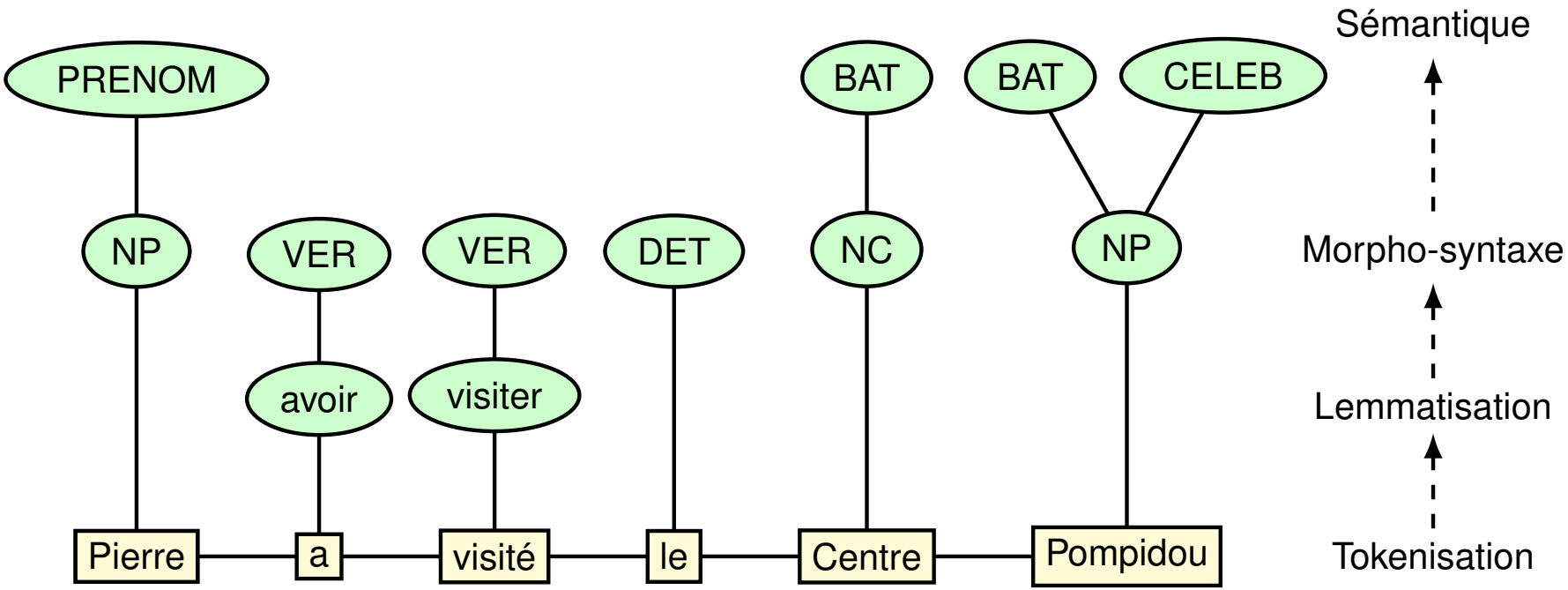
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse :



Traiter automatiquement les langues naturelles

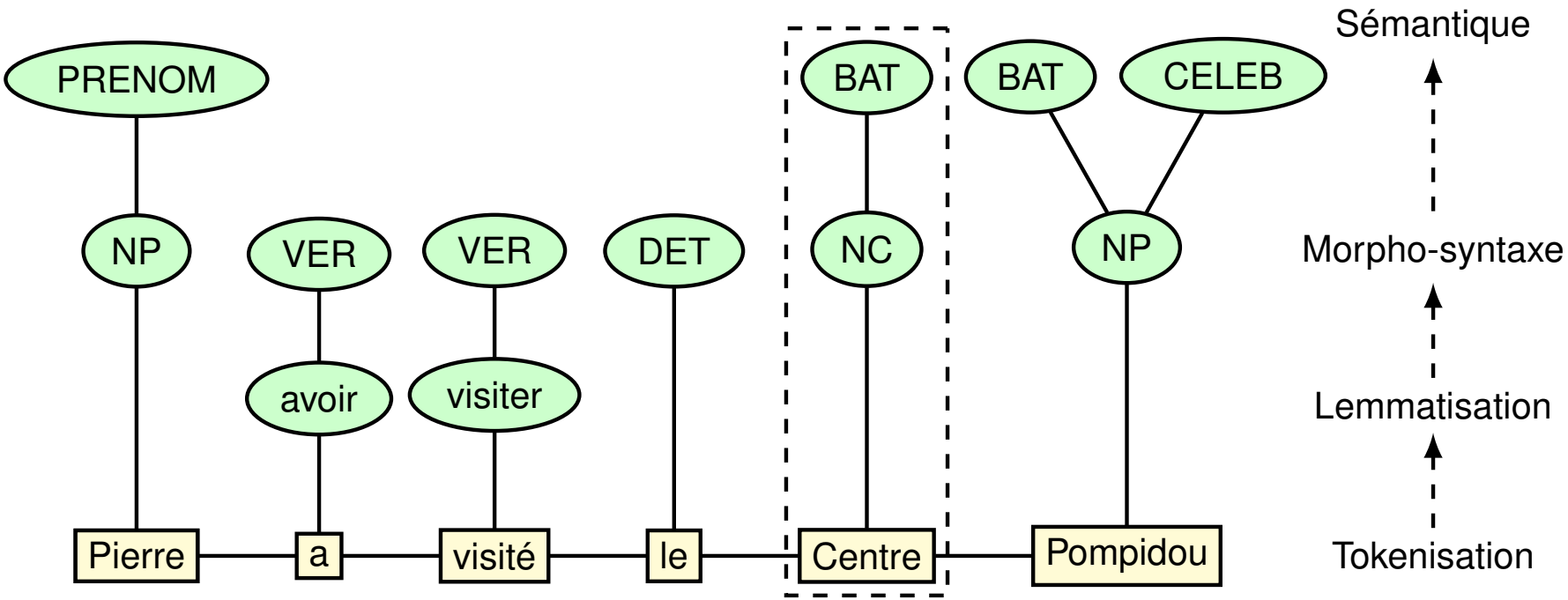
- Stratégie incrémentale d'analyse :



- Axes pour les analyses ultérieures :

Traiter automatiquement les langues naturelles

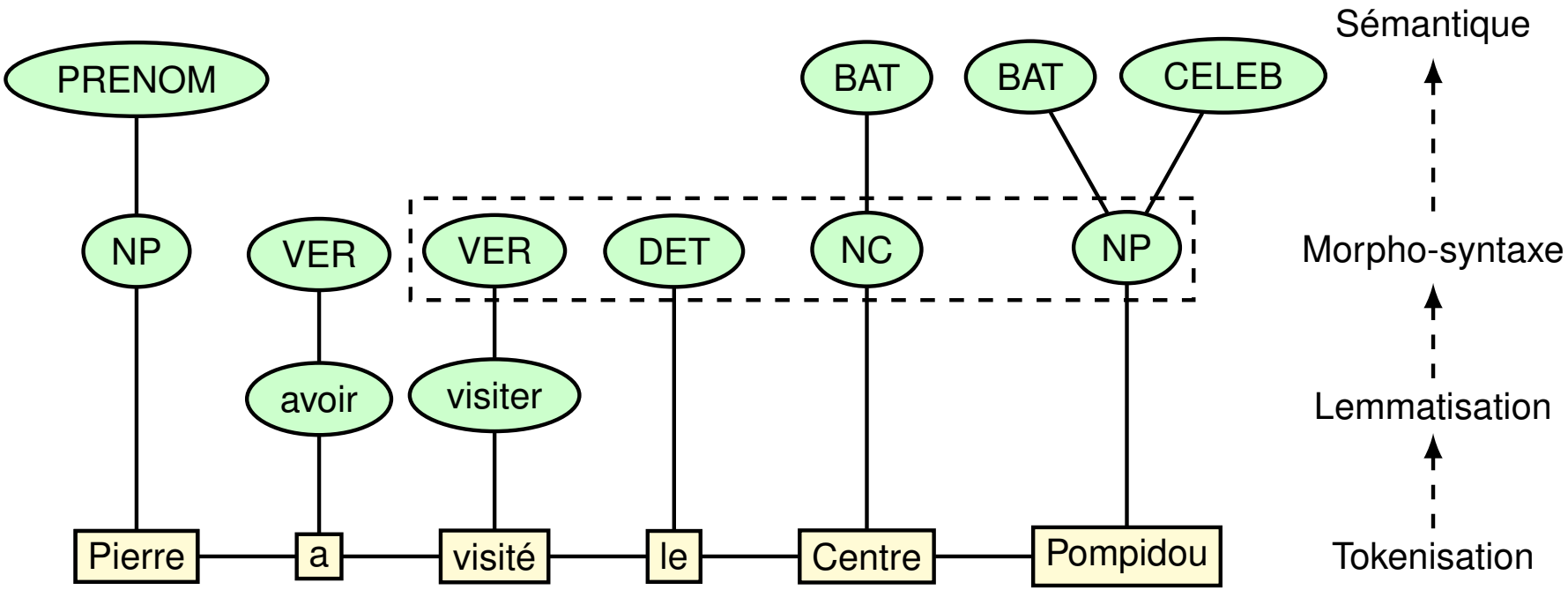
- Stratégie incrémentale d'analyse :



- Axes pour les analyses ultérieures :
  - **Ontologique** : généralisations (formes normales, hyperonymes)

Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse :



- Axes pour les analyses ultérieures :
  - **Ontologique** : généralisations (formes normales, hyperonymes)
  - **Structurel** : prise en compte d'unités contigües (composition)

Approches pour la REN

- Deux grandes classes d'approches :

Approches pour la REN

- Deux grandes classes d'approches :
  - **Orientées connaissances** : des transducteurs (s'appuyant sur les lexiques) sont implémentés **par introspection** selon la connaissance de la problématique

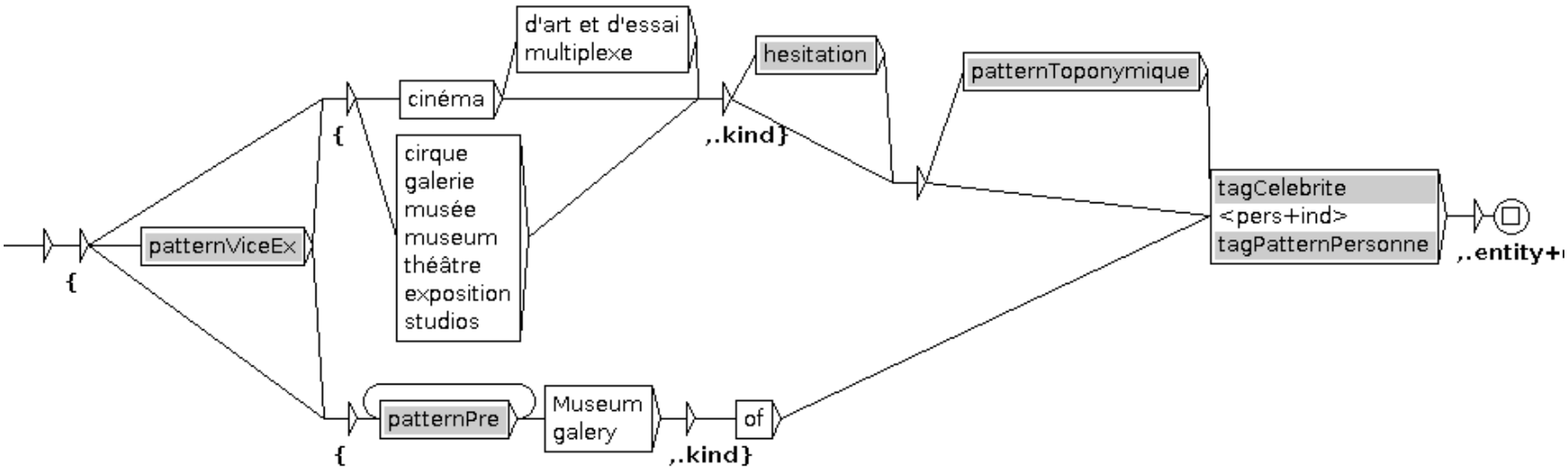
Approches pour la REN

- Deux grandes classes d'approches :
  - **Orientées connaissances** : des transducteurs (s'appuyant sur les lexiques) sont implémentés **par introspection** selon la connaissance de la problématique
  - **Orientées données** : un modèle est **paramétré automatiquement** selon des exemples (corpus) du résultat que l'on souhaite obtenir (fonction objectif)

Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

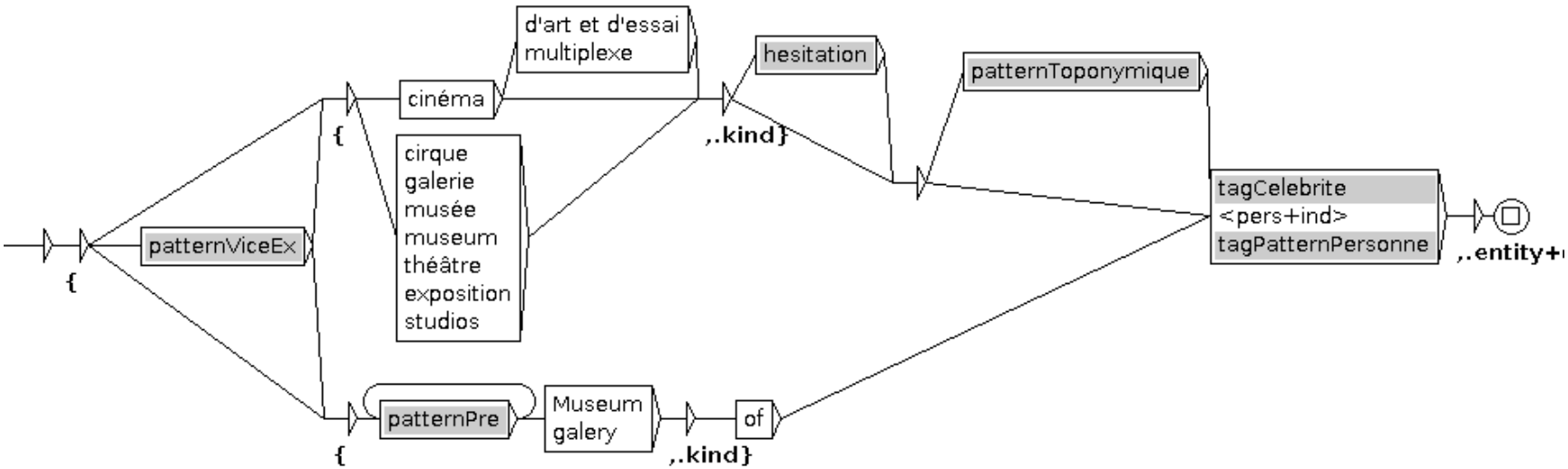
  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



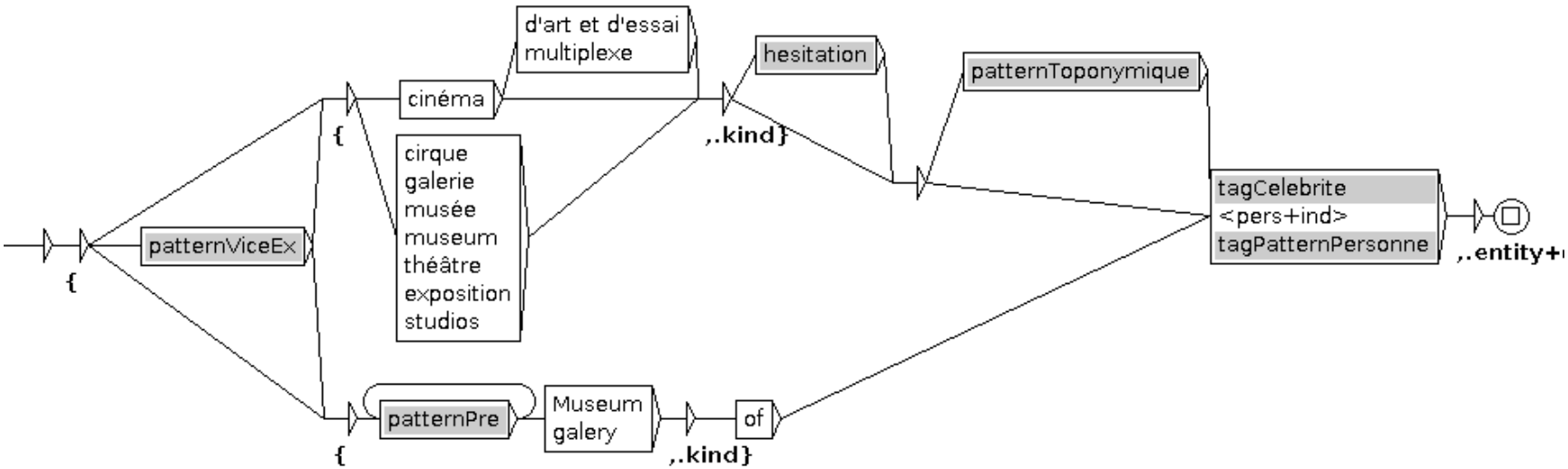
⇒ *+ Très **précise** (contexte) et **contrôlée***

⇒ *+ Connaissance capitalisée*

Approches orientées connaissances

- Approche très répandue à base d'automates ou de transducteurs

  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



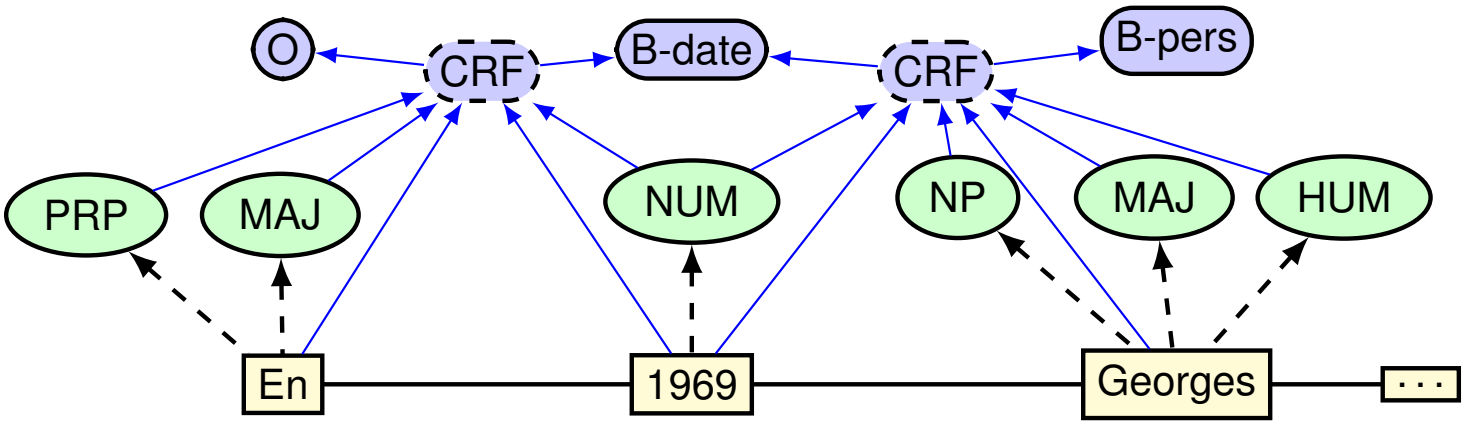
⇒ *+ Très **précise** (contexte) et **contrôlée***

⇒ *+ Connaissance capitalisée*

⇒ *- **Coûteuse** à développer (experts), manque de couverture*

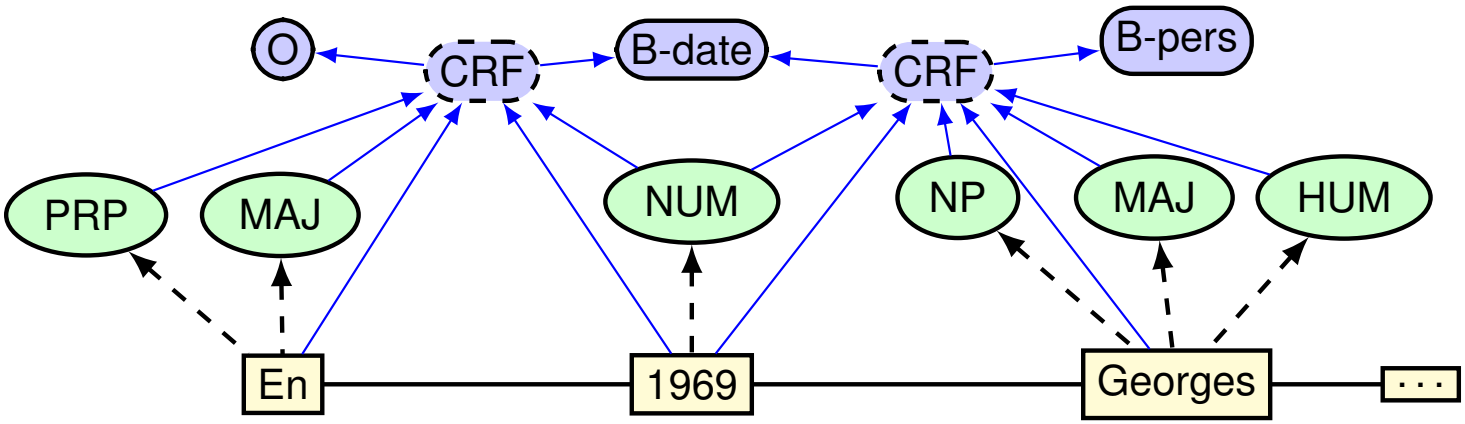
Approches orientées données

- Estimation de la probabilité pour un mot de recevoir un type d'entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010] :
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



Approches orientées données

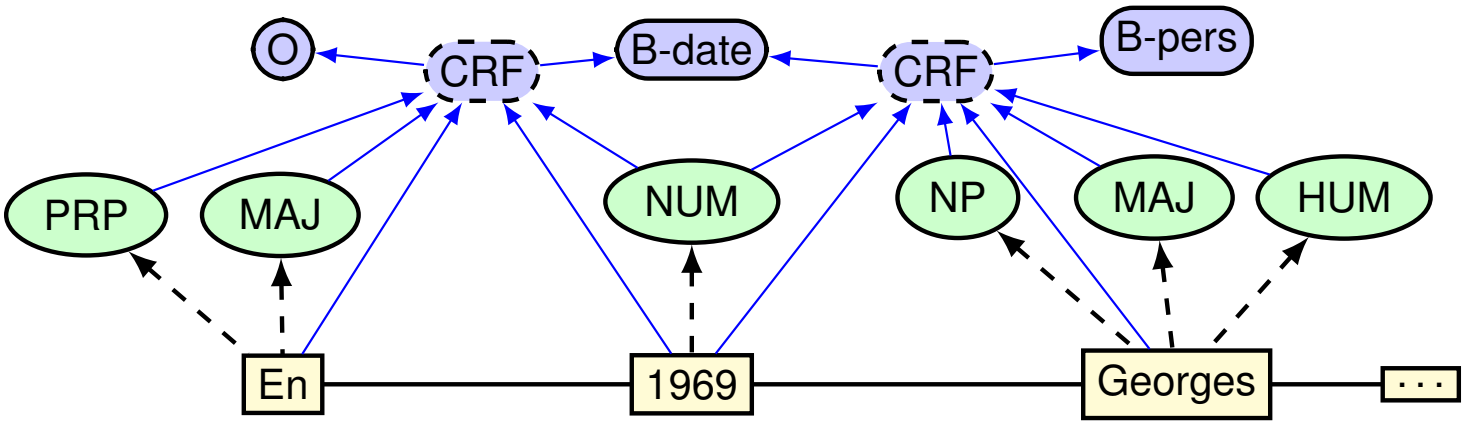
- Estimation de la probabilité pour un mot de recevoir un type d'entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010] :
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



$\Rightarrow$ *+* ***Automatique*** *et* ***robuste***

Approches orientées données

- Estimation de la probabilité pour un mot de recevoir un type d'entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010] :
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



⇒ *+* ***Automatique*** *et* ***robuste***

⇒ *- Décisions locales mot-à-mot*

⇒ *- Difficile à* ***interpréter*** *et à adapter*

Positionnement

Positionnement

- Par rapport aux approches orientées connaissances :
  - $\Rightarrow$ *Extraire des connaissances **complexes** à partir des données*
  - $\Rightarrow$ *Analyses préalables **ambigües***

Positionnement

- Par rapport aux approches orientées connaissances :
  - $\Rightarrow$ *Extraire des connaissances **complexes** à partir des données*
  - $\Rightarrow$ *Analyses préalables **ambigües***

- Par rapport aux approches orientées données :
  - $\Rightarrow$ *Ne pas modéliser un processus "mot-à-mot" ($\neq$ classification)*
  - $\Rightarrow$ *Capacité à générer une **structure** d'annotation*

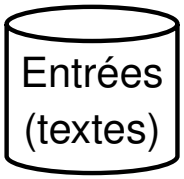
Positionnement

- Par rapport aux approches orientées connaissances :
  - $\Rightarrow$ *Extraire des connaissances **complexes** à partir des données*
  - $\Rightarrow$ *Analyses préalables **ambigües***

- Par rapport aux approches orientées données :
  - $\Rightarrow$ *Ne pas modéliser un processus "mot-à-mot" ($\neq$ classification)*
  - $\Rightarrow$ *Capacité à générer une **structure** d'annotation*

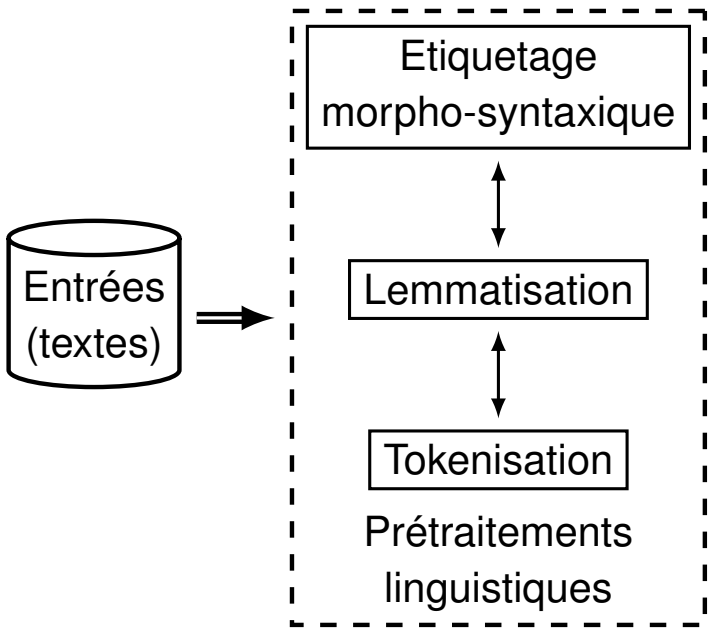
$\Rightarrow$ *Originalité : rechercher séparément les **marqueurs** qui délimitent le **début** ou la **fin** d'une entité nommée*

Cadre expérimental



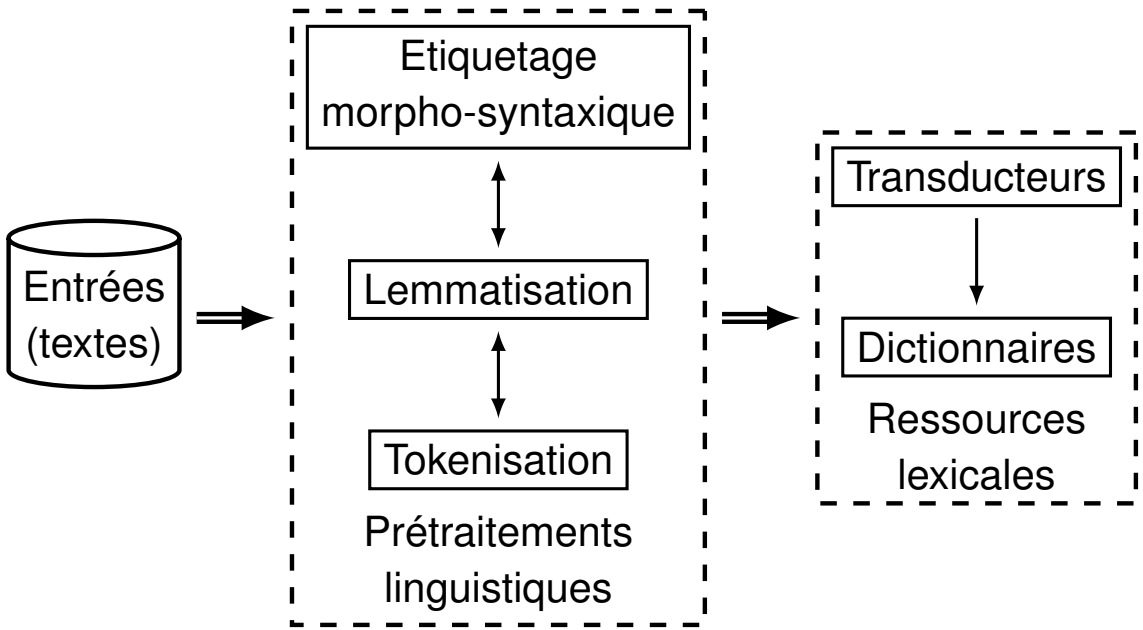
Cadre expérimental

- Prétraitements : **TreeTagger** avec adaptations (nombres, noms propres, etc.)



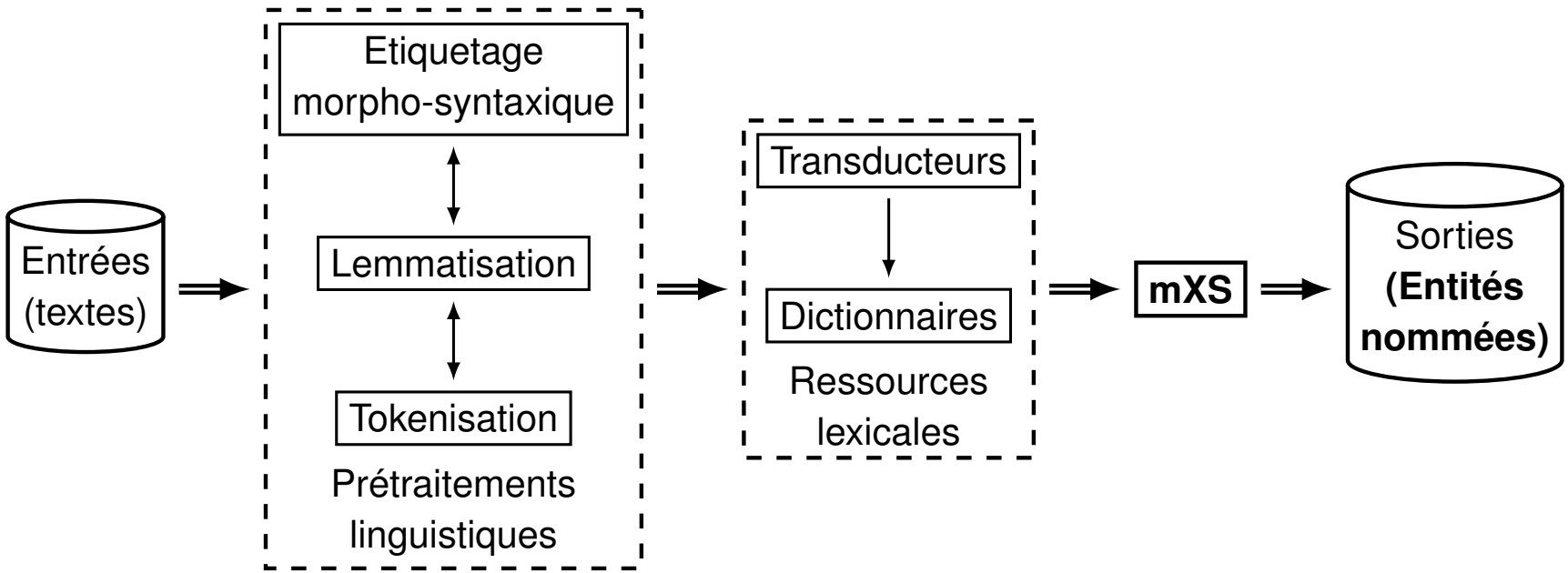
Cadre expérimental

- Prétraitements : **TreeTagger** avec adaptations (nombres, noms propres, etc.)
- Lexiques : **CasEN** et manuels avec adaptations (159 catégories, 221 547 mots, 443 112 catégorisations)



Cadre expérimental

- Prétraitements : **TreeTagger** avec adaptations (nombres, noms propres, etc.)
- Lexiques : **CasEN** et manuels avec adaptations (159 catégories, 221 547 mots, 443 112 catégorisations)



Plan

Fouille de données textuelles

Fouille de données textuelles

- Explorer des données pour **enrichir des lexiques** :
  - A partir de corpus [Riloff & Jones 1999]
  - A partir du web [Mooney & Bunescu 2005, Nadeau 2007, Béchet & Roche 2010]
  - A partir d'encyclopédies (Wikipedia, geonames) [Bunescu Pasca 2006, Charton 2009]

Fouille de données textuelles

- Explorer des données pour **enrichir des lexiques** :
  - A partir de corpus [Riloff & Jones 1999]
  - A partir du web [Mooney & Bunescu 2005, Nadeau 2007, Béchet & Roche 2010]
  - A partir d'encyclopédies (Wikipedia, geonames) [Bunescu Pasca 2006, Charton 2009]
- Explorer des données pour **extraire des motifs** :
  - Automates [Hingston 2002, Mendes & Antunes 2009]
  - Grammaires locales dédiées [Besancon et.al. 2006, Sun & Grishman 2010]
  - Motifs séquentiels [Plantevit et. al. 2009, Cellier & Charnois 2010]
  - Extracteurs d'entités nommées [Kushmerick 1997, Califf & Mooney 1999]

Fouille de données textuelles

- Explorer des données pour **enrichir des lexiques** :
  - A partir de corpus [Riloff & Jones 1999]
  - A partir du web [Mooney & Bunescu 2005, Nadeau 2007, Béchet & Roche 2010]
  - A partir d'encyclopédies (Wikipedia, geonames) [Bunescu Pasca 2006, Charton 2009]
- Explorer des données pour **extraire des motifs** :
  - Automates [Hingston 2002, Mendes & Antunes 2009]
  - Grammaires locales dédiées [Besancon et.al. 2006, Sun & Grishman 2010]
  - Motifs séquentiels [Plantevit et. al. 2009, Cellier & Charnois 2010]
  - Extracteurs d'entités nommées [Kushmerick 1997, Califf & Mooney 1999]

$\Rightarrow$ *Extraction de **motifs séquentiels hiérarchiques** [EGC'10,LTC'11] à partir de corpus volumineux*

Exploration de données enrichies

- **Base de données** : données (séquences de tokens, corpus)

Items, séquences et motifs

$E =$ `'Il visite Washington .'`

Exploration de données enrichies

- **Base de données** : données (séquences de tokens, corpus)
- Enrichissement des données :
  - Application $R()$ sur des **séquences d'items**

Items, séquences et motifs

$E =$ 'Il visite Washington .'

$R(E) =$ 'PRO/il/Il VER/visiter/visite
VILLE/NP/Washington⊕CELEB/NP/Washington PONCT/.'

Exploration de données enrichies

- **Base de données** : données (séquences de tokens, corpus)
- Enrichissement des données :
  - Application $R()$ sur des **séquences d'items**

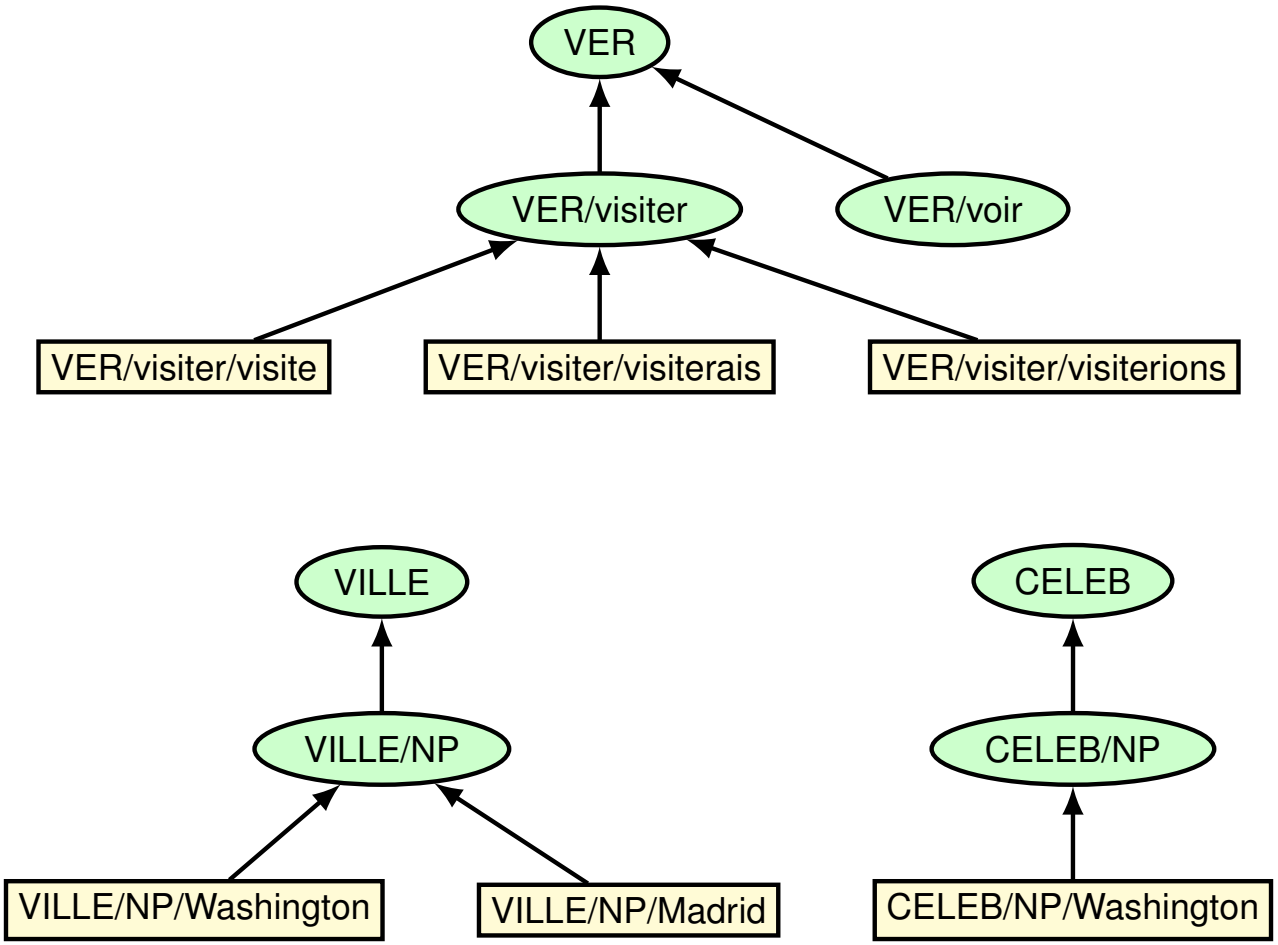
Items, séquences et motifs

$E =$ 'Il visite Washington .'

$R(E) =$ 'PRO/il/Il VER/visiter/visite VILLE/NP/Washington⊕CELEB/NP/Washington PONCT/.'

- Formalisation des motifs
  - **Alphabet des motifs** : symboles des données enrichies (analyses),
    $\Rightarrow$ *Généralisation des items au sein d'une* ***hiérarchie***
  - **Motifs** : par concaténation sur l'alphabet de motifs,
  - **Couverture et mesures** : mesures de l'*intérêt* des motifs

Hiérarchies d'items



Généralisation de motifs par la hiérarchie

- **Définition** : soient deux motifs $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ et $Q = q_1 q_2 \dots q_n \in \mathcal{L}_p$, alors P *généralise hiérarchiquement* Q, noté $Q \leq_h P$, si pour tout $j \in [1, n]$, alors $q_j \leq_h p_j$.

$\Rightarrow$ *Formes normales et hyperonymes au sein des motifs*

Exemple de généralisation

'A B/D C/E' $\leq_h$ 'A B/D C' $\leq_h$ 'A B C'

Exemple lié aux entités nommées

'PRO VERB/visiter VILLE/NP/Madrid'
$\leq_h$ 'PRO VERB/visiter VILLE'

Généralisation de motifs par affixation

- **Définition** : soient deux motifs $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ et $Q = q_1 q_2 \dots q_p \in \mathcal{L}_p$, alors P *généralise par affixation* Q, noté $Q \leq_a P$, si $p \geq n$ et s'il existe au moins un $k \in [0, p-n]$ tel que, pour tout $j \in [1, n]$, alors $q_{j+k} = p_j$.

$\Rightarrow$ *Extraire des portions d'énoncés comme motifs*

Exemple de généralisation

'A B C' $\leq_a$ 'B C' $\leq_a$ 'B'

Exemple lié aux entités nommées

'PRO VERB/visiter VILLE/NP' $\leq_a$ 'VERB/visiter VILLE/NP'

Généralisation de motifs sur marqueurs

- **Définition** : soient deux motifs $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ et $Q = q_1 q_2 \dots q_p \in \mathcal{L}_p$, alors P *généralise sur marqueurs* Q, noté $Q \leq_m P$, si $p \geq n$ et s'il existe une fonction *discrète strictement croissante* $C()$ définie de $[1, n]$ vers $[1, p]$ telle que, pour tout $j \in [1, n]$, alors $q_{C(j)} = p_j$ et, pour tout $k \in [1, p]$ tel que $k \notin \{C(j), j \in [1, n]\}$, alors $q_k \in \Sigma_m$.

$\Rightarrow$ *Mesurer la corrélation des motifs aux marqueurs*

Exemple de généralisation

'<pers> A </pers> <loc> B </loc>' $\leq_m$ 'A <loc> B' $\leq_m$ 'A B'

Exemple lié aux entités nommées

'PRO VERB/visiter <loc> NP </loc>' $\leq_m$ 'PRO VERB/visiter NP'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

`'VERB/visiter VILLE/NP'`

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

'VERB/visiter VILLE/NP'

$\Rightarrow$ 'VERB/visiter <loc> VILLE/NP </loc>'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

'VERB/visiter VILLE/NP'

$\Rightarrow$ 'VERB/visiter <loc> VILLE/NP </loc>'

'VERB/rencontrer CELEB/NP'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

'VERB/visiter VILLE/NP'

$\Rightarrow$ 'VERB/visiter <loc> VILLE/NP </loc>'

'VERB/rencontrer CELEB/NP'

$\Rightarrow$ 'VERB/rencontrer <pers> CELEB/NP'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

'VERB/visiter VILLE/NP'

$\Rightarrow$ 'VERB/visiter <loc> VILLE/NP </loc>'

'VERB/rencontrer CELEB/NP'

$\Rightarrow$ 'VERB/rencontrer <pers> CELEB/NP'

'DET/Le/le NUM juillet NUM PONCT/,'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

'VERB/visiter VILLE/NP'

$\Rightarrow$ 'VERB/visiter <loc> VILLE/NP </loc>'

'VERB/rencontrer CELEB/NP'

$\Rightarrow$ 'VERB/rencontrer <pers> CELEB/NP'

'DET/Le/le NUM juillet NUM PONCT/,'

$\Rightarrow$ 'DET/Le/le <date> NUM juillet NUM </date> PONCT/,'

Les règles d'annotation

- Reconnaître les entités nommées $\approx$ insérer des **marqueurs** (transduction, annotation)
- **Règle d'annotation** : une règle d'annotation R est un motif $P = p_1 p_2 \ldots p_n \in \mathcal{L}_p$ tel qu'il existe au moins un j pour lequel $p_j \in \Sigma_m$ et un k pour lequel $p_k \notin \Sigma_m$.

Exemples de règles d'annotation

```
'VERB/visiter VILLE/NP'
    ⇒ 'VERB/visiter <loc> VILLE/NP </loc>'
'VERB/rencontrer CELEB/NP'
    ⇒ 'VERB/rencontrer <pers> CELEB/NP'
'DET/Le/le NUM juillet NUM PONCT/,'
    ⇒ 'DET/Le/le <date> NUM juillet NUM </date> PONCT/,'
```

- Pas d'items **lexicaux** (instances de noms propres, flexions)

Critères pour extraire les règles

Critères pour extraire les règles

- $\Rightarrow$ ***Filtrer*** *les règles qui ont a priori peu d'intérêt*
- $\Rightarrow$ *Nécessité de comptabiliser les occurrences* $Occ(P, R(E))$

Critères pour extraire les règles

$\Rightarrow$ ***Filtrer*** *les règles qui ont a priori peu d'intérêt*

$\Rightarrow$ *Nécessité de comptabiliser les occurrences* $Occ(P, R(E))$

- Règles **fréquentes** :
  - Suffisamment générales pour être productives

$$Freq(P, \mathcal{D}) = \sum_{E \in \mathcal{D}} |Occ(P, R(E))|$$

Critères pour extraire les règles

$\Rightarrow$ ***Filtrer*** *les règles qui ont a priori peu d'intérêt*

$\Rightarrow$ *Nécessité de comptabiliser les occurrences* $Occ(P, R(E))$

- Règles **fréquentes** :
  - Suffisamment générales pour être productives

$$Freq(P, \mathcal{D}) = \sum_{E \in \mathcal{D}} |Occ(P, R(E))|$$

- Règles **confiantes** :
  - Suffisamment précises pour indiquer des annotations correctes

$$Conf(P, \mathcal{D}) = \frac{Freq(P, \mathcal{D})}{Freq(Ret_m(P), \mathcal{D})}$$

Critères pour extraire les règles

$\Rightarrow$ ***Filtrer*** *les règles qui ont a priori peu d'intérêt*

$\Rightarrow$ *Nécessité de comptabiliser les occurrences* $Occ(P, R(E))$

- Règles **fréquentes** :
  - Suffisamment générales pour être productives

$$Freq(P, \mathcal{D}) = \sum_{E \in \mathcal{D}} |Occ(P, R(E))|$$

- Règles **confiantes** :
  - Suffisamment précises pour indiquer des annotations correctes

$$Conf(P, \mathcal{D}) = \frac{Freq(P, \mathcal{D})}{Freq(Ret_m(P), \mathcal{D})}$$

- Règles **informatives** :
  - Deux règles de même fréquence liées par généralisation

    $\Rightarrow$ *Couvrent les mêmes exemples*

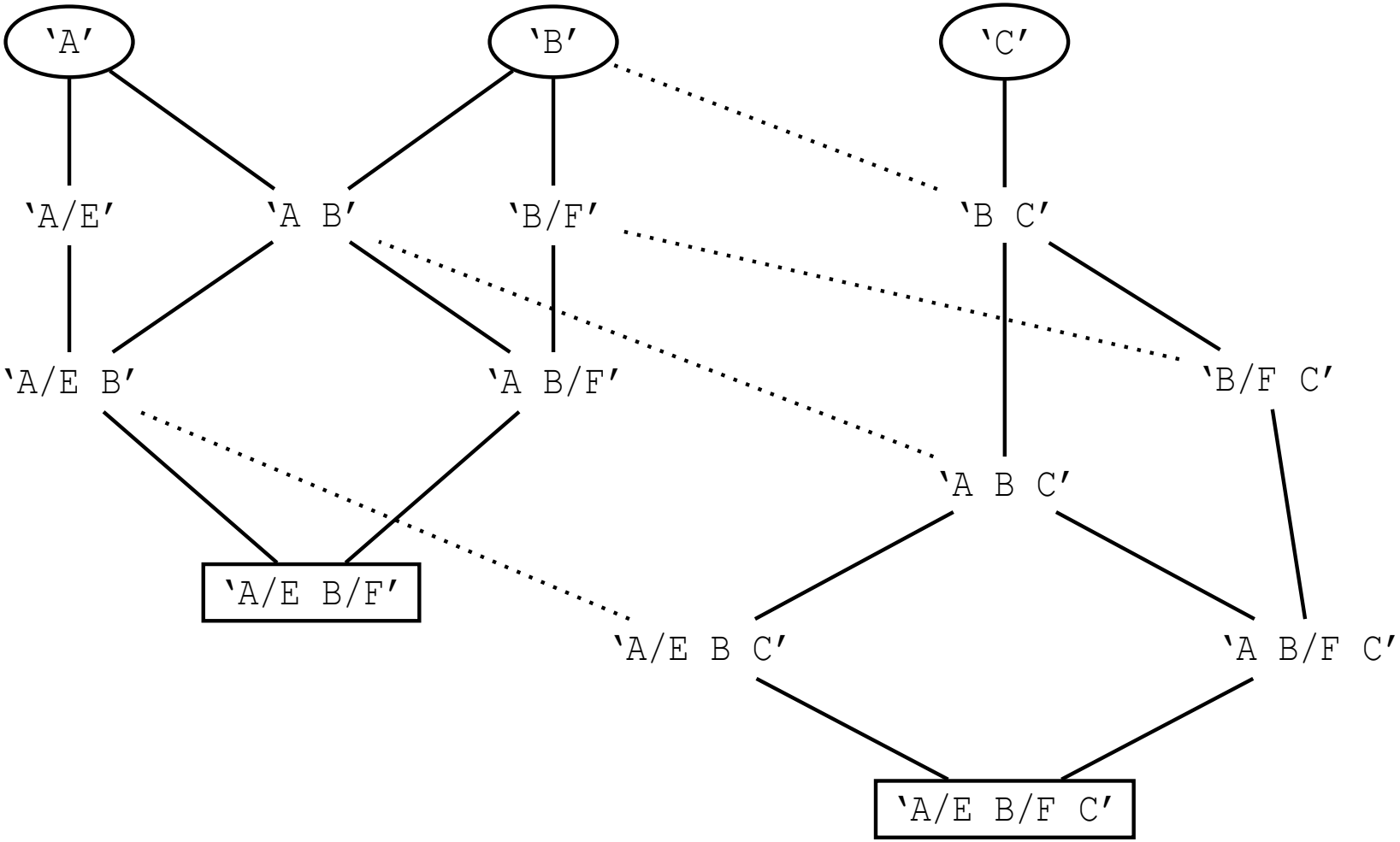
Hiérarchie des motifs

Exemples de base de données

$R(E_1) =$ 'A/E B/F C/G'
$R(E_2) =$ 'A/E B/F C/H'
$R(E_3) =$ 'A/E B/F D/G'

	$n=1$	$n=2$	$n=3$
$f=2$	'C'	'B C' 'B/F C'	'A B C' 'A/E B C' 'A B/F C' 'A/E B/F C'
$f=3$	'A' 'A/E' 'B' 'B/F'	'A B' 'A/E B' 'A B/F' 'A/E B/F'	

Hiérarchie des motifs



Combinatoire des motifs

- Forte combinatoire lorsque des items se répètent

Combinatoire des motifs

- Forte combinatoire lorsque des items se répètent

Exemple de combinatoire des motifs

'<pers> Dali </pers>'

`'<pers> CELEB/NP </pers>'`

`'<pers> CELEB </pers>'`

'<pers> Georges Pompidou </pers>'

`'<pers> CELEB/NP CELEB/NP </pers>'`

`'<pers> CELEB CELEB/NP </pers>'`

`'<pers> CELEB/NP CELEB </pers>'`

`'<pers> CELEB CELEB </pers>'`

'<pers> Valery Giscard d'Estaing </pers>'

`'<pers> CELEB/NP CELEB/NP CELEB/DET/de CELEB/NP </pers>'`

. . . (total : 26 motifs)

$\Rightarrow$ *Proposition : un item couvre un **segment** de données*

Proposition : motifs de segments

- **Motif de segments** : un motif de segments est un motif $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ tel que, pour tout $j \in [1, n-1]$ et $k \in [j+1, n]$ tels que tout $l \in [j+1, l-1]$ vérifie $p_l \in \Sigma_m$, alors $p_j \not\trianglelefteq_h p_k$ et $p_l \not\trianglelefteq_h p_j$.

  $\Rightarrow$ *Contrainte sur les items contigus (anti-monotonie)*

  $\Rightarrow$ *Redéfinir la* ***couverture*** *et les* ***généralisations***

Proposition : motifs de segments

- **Motif de segments** : un motif de segments est un motif $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ tel que, pour tout $j \in [1, n-1]$ et $k \in [j+1, n]$ tels que tout $l \in [j+1, l-1]$ vérifie $p_l \in \Sigma_m$, alors $p_j \not\leq_h p_k$ et $p_l \not\leq_h p_j$.

  $\Rightarrow$ *Contrainte sur les items contigus (anti-monotonie)*

  $\Rightarrow$ *Redéfinir la* ***couverture*** *et les* ***généralisations***

Réduction de la combinatoire des motifs

'<pers> Valery Giscard d'Estaing </pers>'

`'<pers> CELEB/NP CELEB/PREP/de CELEB/NP </pers>'`

`'<pers> CELEB/NP CELEB/PREP CELEB/NP </pers>'`

`'<pers> CELEB </pers>'`

Proposition : motifs de segments

- **Motif de segments** : un motif de segments est un motif $P = p_1 p_2 \dots p_n \in \mathcal{L}_p$ tel que, pour tout $j \in [1, n-1]$ et $k \in [j+1, n]$ tels que tout $l \in [j+1, l-1]$ vérifie $p_l \in \Sigma_m$, alors $p_j \not\leq_h p_k$ et $p_l \not\leq_h p_j$.

  $\Rightarrow$ *Contrainte sur les items contigus (anti-monotonie)*

  $\Rightarrow$ *Redéfinir la* ***couverture*** *et les* ***généralisations***

Réduction de la combinatoire des motifs

'<pers> Valery Giscard d'Estaing </pers>'

`'<pers> CELEB/NP CELEB/PREP/de CELEB/NP </pers>'`

`'<pers> CELEB/NP CELEB/PREP CELEB/NP </pers>'`

`'<pers> CELEB </pers>'`

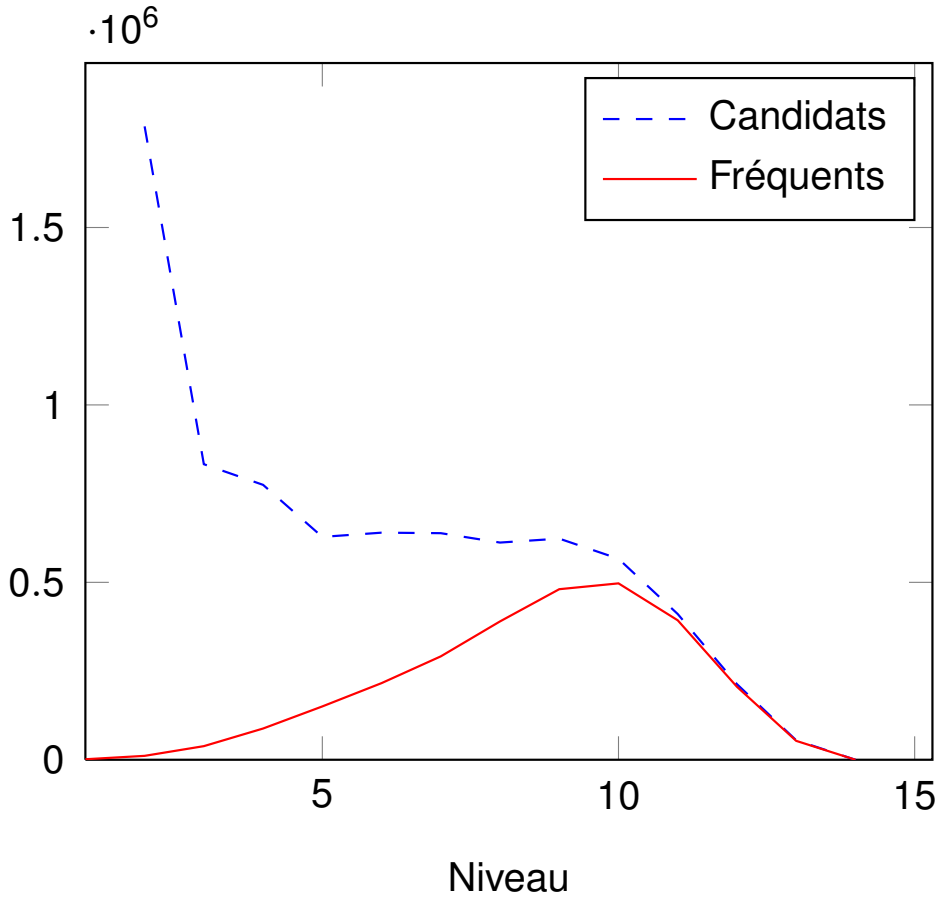
$\Rightarrow$ *Motifs* ***plus courts*** *et plus* ***fréquents*** *(couvrent plus d'occurrences)*

Extraction des motifs

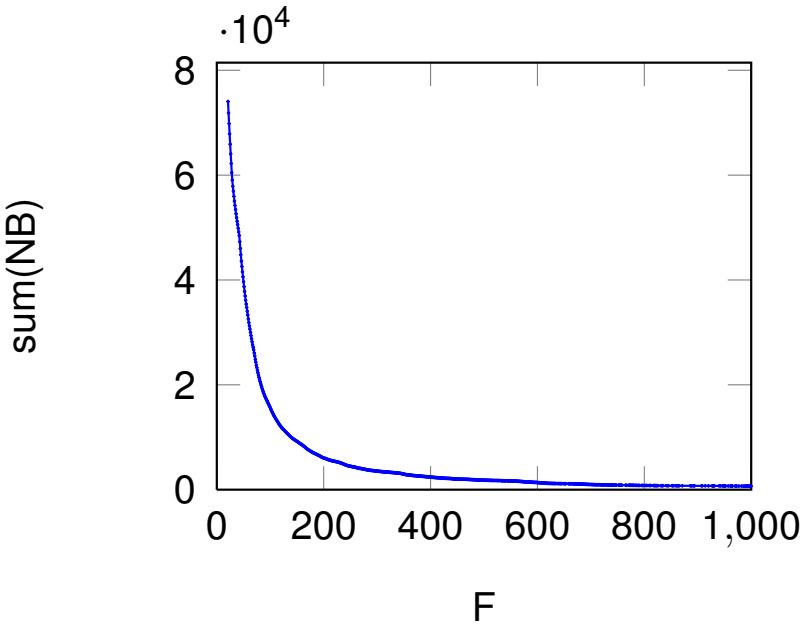
- Algorithme par niveaux (avec optimisations) :

Extraction des motifs

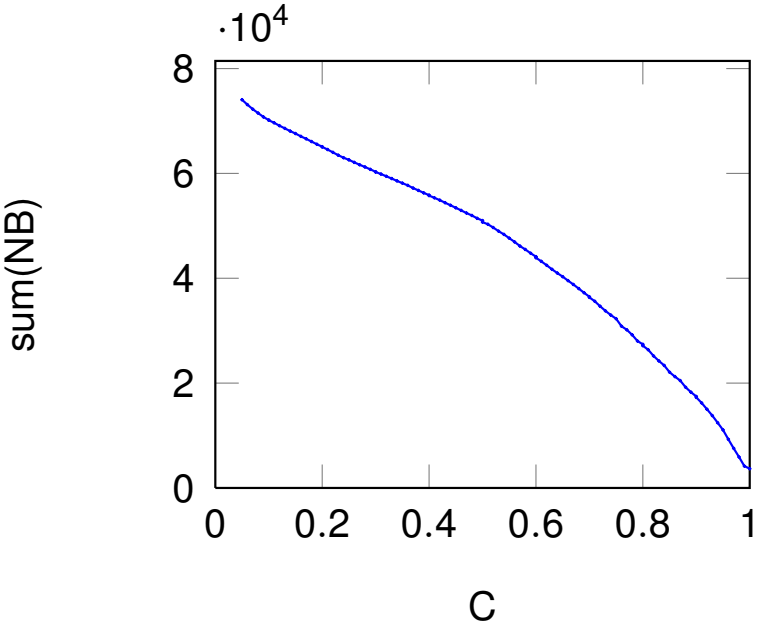
- Algorithme par niveaux (avec optimisations) :



Nombre de motifs extraits

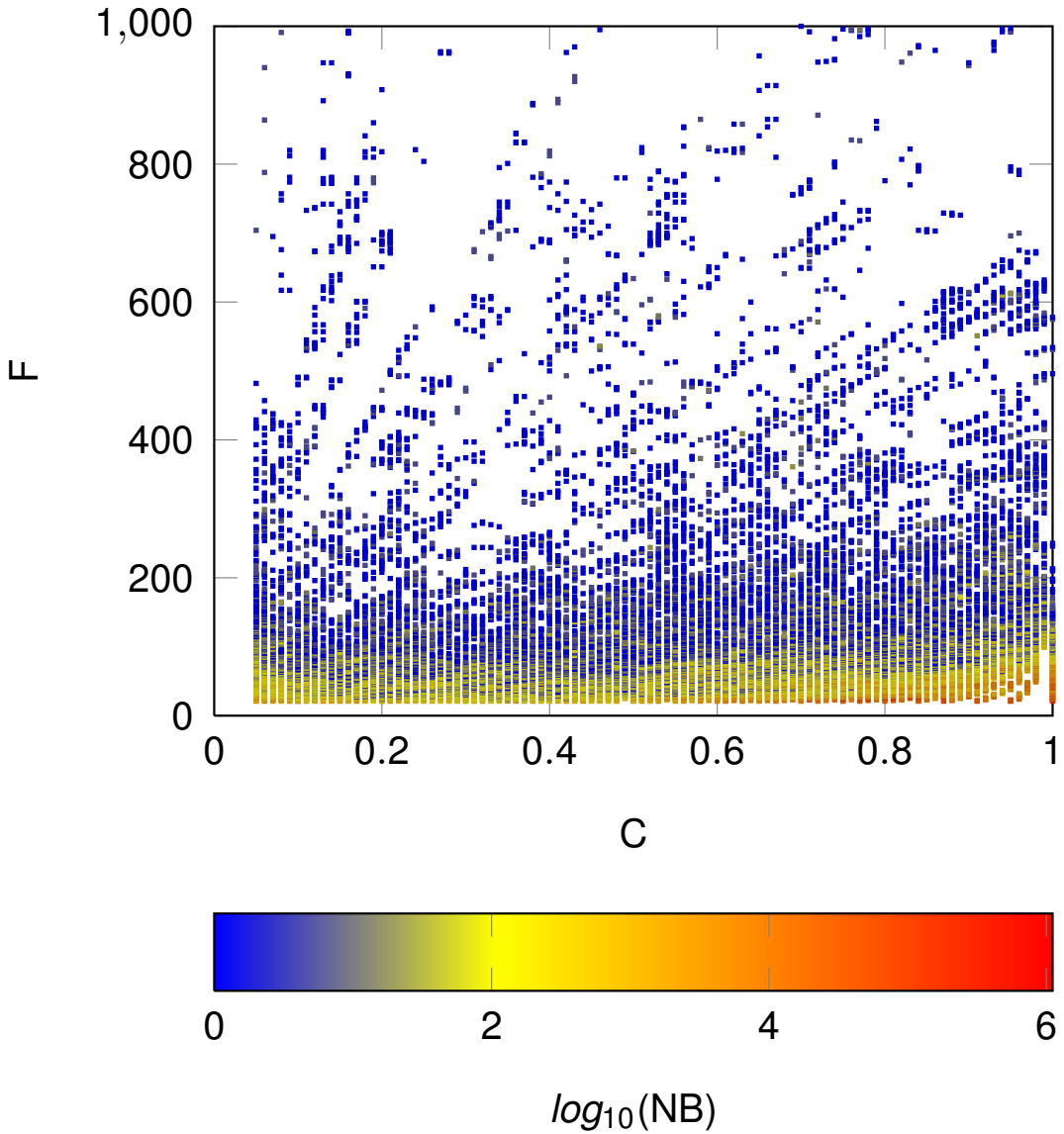


(g) Fréquence



(h) Confiance

Spectre des motifs extraits



Plan

Prédiction des marqueurs (Règles, Bayes)

- Appliquer les règles (Règles) :
  - **Complètes** : doivent réaliser une annotation valide
  - Ordonnées selon leur confiance

  $\Rightarrow$ *Ajout de marqueurs tant qu'il reste des règles à appliquer*

Prédiction des marqueurs (Règles, Bayes)

- Appliquer les règles (Règles) :
  - **Complètes** : doivent réaliser une annotation valide
  - Ordonnées selon leur confiance

  $\Rightarrow$ *Ajout de marqueurs tant qu'il reste des règles à appliquer*
- Combiner les règles (Bayes) :
  - **Modèle probabiliste sur les marqueurs**
  - Statistiques du corpus pour estimer la probabilité d'un marqueur :

$$P(m \in M_j | \mathcal{R}_j) \approx P(m) * \frac{\prod_{R \in \mathcal{R}_j} P(R|m)}{\prod_{R \in \mathcal{R}_j} P(R)}$$

$\Rightarrow$ *Calcul de l'annotation la plus probable :*

$$P(M_1 M_2 \dots M_n) \approx \prod_{j=1}^{n} P(M_j)$$

Modèles de marqueurs - Logit

- Pondération de règles :
  - Modèle probabiliste sur marqueurs (régression logistique, Scikit [Pedregosa et. al. 2012])
  - **Paramétrage itératif orienté données**

Modèles de marqueurs - Logit

- Pondération de règles :
  - Modèle probabiliste sur marqueurs (régression logistique, Scikit [Pedregosa et. al. 2012])
  - **Paramétrage itératif orienté données**
- Formulation : $P(m \in M_j | \mathcal{R}_j) = \frac{exp\left(\sum_{R \in \mathcal{R}_j} \lambda_{R,m}\right)}{Z(\mathcal{R}_j)}$

$\Rightarrow$ *Calcul des marqueurs les plus probables*

Modèles de marqueurs - Logit

- Pondération de règles :
  - Modèle probabiliste sur marqueurs (régression logistique, Scikit [Pedregosa et. al. 2012])
  - **Paramétrage itératif orienté données**
- Formulation : $P(m \in M_j | \mathcal{R}_j) = \frac{exp\left(\sum_{R \in \mathcal{R}_j} \lambda_{R,m}\right)}{Z(\mathcal{R}_j)}$

$\Rightarrow$ *Calcul des marqueurs les plus probables*

- Annotation structurée : séquences de marqueurs

$$P(M_j = m_1 m_2 \dots m_p) = \frac{\sum_{k=1}^{p} P(m_k \in M_k | \mathcal{R}_k) P(m_1 \dots m_p | m_k)}{p}$$

$\Rightarrow$ *Calcul de l'annotation la plus probable :*

$$P(M_1 M_2 \dots M_n) \approx \prod_{j=1}^{n} P(M_j)$$

Exemple d'application

Données

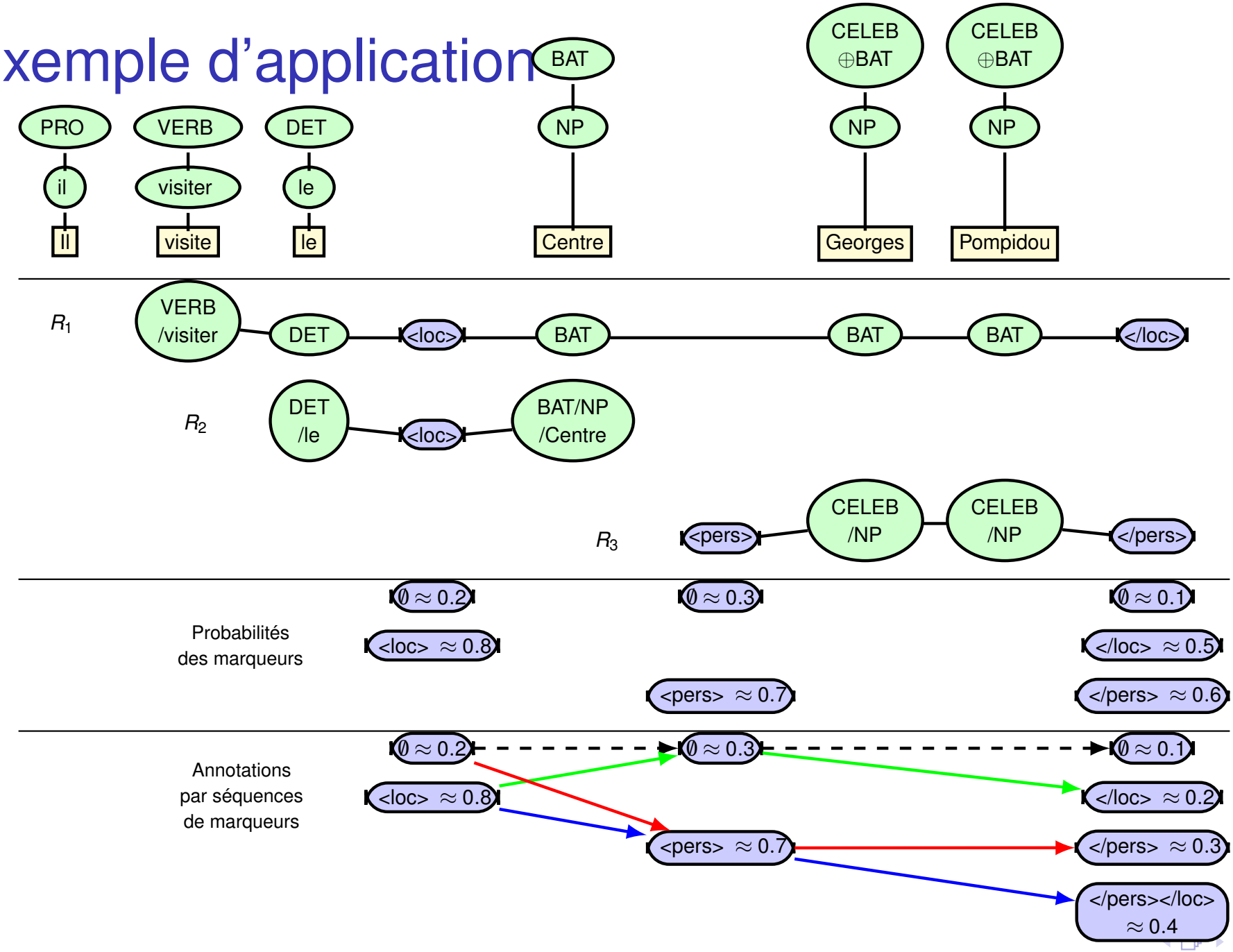
'Il visite le Centre Georges Pompidou'

R_1 : 'VERB/visiter DET <loc> BAT </loc>'

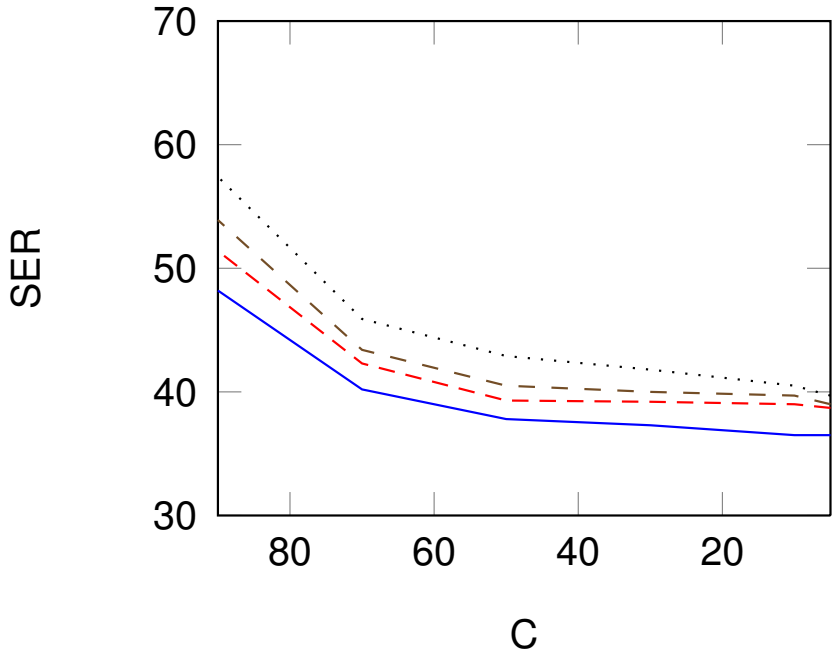
R_2 : 'DET/le <loc> BAT/NP/Centre'

R_3 : '<pers> CELEB/NP </pers>'

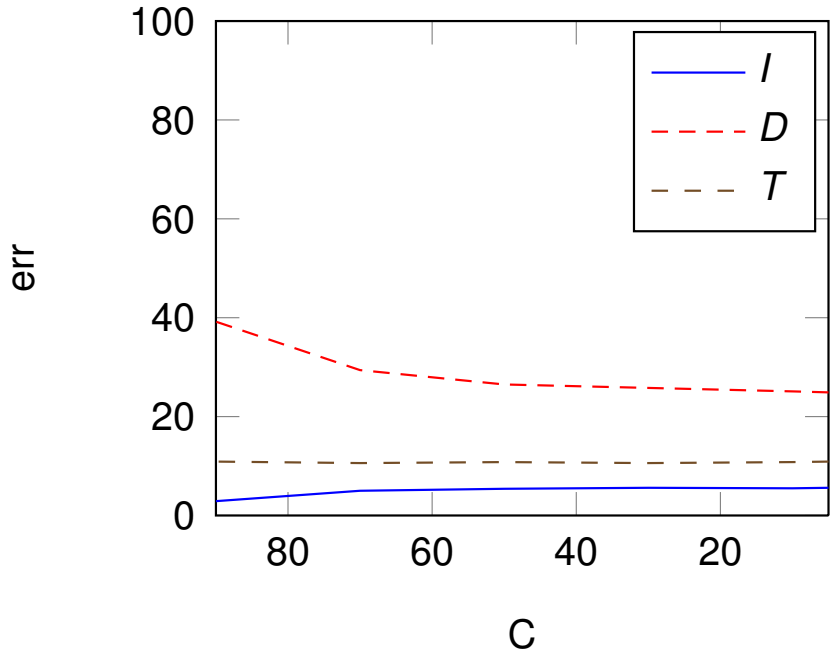
Exemple d'application



Performances globales



(i) Performances

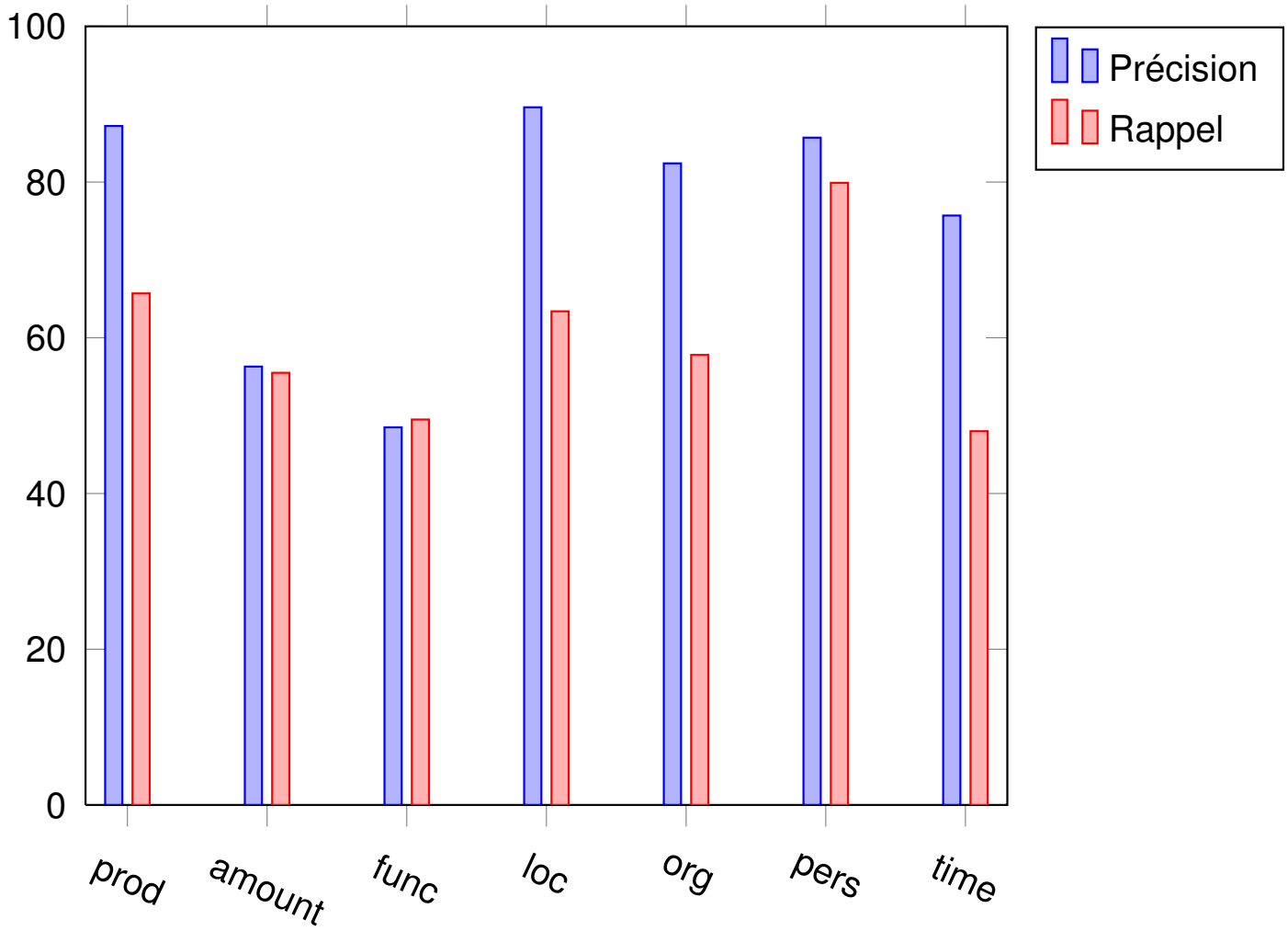


(j) Erreurs, $F \geq 5.10^{-5}$

SER : taux d'erreur (à minimiser), erreurs de délétion (D), d'insertion (I) et de substitution (S) :

$$SER(ref, hyp)) = \frac{|D| + |I| + |S|}{|Slots(ref)|}$$

Performances par type



Performances selon la configuration

Approche	F	C	Règles	SER	I	D	S	P	R	Fm
Logit	03	10	106 062	**36,1**	5,7	24,4	10,9	79,6	64,7	71,4
Logit+Segs	03	05	143 205	**35,9**	5,6	24,2	10,8	79,8	64,9	71,6
Logit-Dicos	03	05	80 231	**45,2**	5,9	30,2	16,3	70,7	53,5	60,9
Logit+Test	03	05	141 550	**26,3**	3,2	18,6	8,1	86,6	73,3	79,4
Logit-D25	03	05	100 027	**36,2**	5,6	24,6	10,9	79,7	64,6	71,3
Logit-D50	03	05	73 332	**36,7**	5,4	25,2	11,0	79,5	63,8	70,8
Logit-D75	03	05	50 408	**39,0**	5,4	27,0	11,7	78,2	61,3	68,7

P : précision, **R** : rappel, **Fm** : F-mesure

Performances selon la configuration

Approche	F	C	Règles	SER	I	D	S	P	R	Fm
Logit	03	10	106 062	**36,1**	5,7	24,4	10,9	79,6	64,7	71,4
Logit+Segs	03	05	143 205	**35,9**	5,6	24,2	10,8	79,8	64,9	71,6
Logit-Dicos	03	05	80 231	**45,2**	5,9	30,2	16,3	70,7	53,5	60,9
Logit+Test	03	05	141 550	**26,3**	3,2	18,6	8,1	86,6	73,3	79,4
Logit-D25	03	05	100 027	**36,2**	5,6	24,6	10,9	79,7	64,6	71,3
Logit-D50	03	05	73 332	**36,7**	5,4	25,2	11,0	79,5	63,8	70,8
Logit-D75	03	05	50 408	**39,0**	5,4	27,0	11,7	78,2	61,3	68,7

P : précision, **R** : rappel, **Fm** : F-mesure

Performances selon la configuration

Approche	F	C	Règles	SER	I	D	S	P	R	Fm
Logit	03	10	106 062	**36,1**	5,7	24,4	10,9	79,6	64,7	71,4
Logit+Segs	03	05	143 205	**35,9**	5,6	24,2	10,8	79,8	64,9	71,6
Logit-Dicos	03	05	80 231	**45,2**	5,9	30,2	16,3	70,7	53,5	60,9
Logit+Test	03	05	141 550	**26,3**	3,2	18,6	8,1	86,6	73,3	79,4
Logit-D25	03	05	100 027	**36,2**	5,6	24,6	10,9	79,7	64,6	71,3
Logit-D50	03	05	73 332	**36,7**	5,4	25,2	11,0	79,5	63,8	70,8
Logit-D75	03	05	50 408	**39,0**	5,4	27,0	11,7	78,2	61,3	68,7

P : précision, **R** : rappel, **Fm** : F-mesure

Performances selon la configuration

Approche	F	C	Règles	SER	I	D	S	P	R	Fm
Logit	03	10	106 062	**36,1**	5,7	24,4	10,9	79,6	64,7	71,4
Logit+Segs	03	05	143 205	**35,9**	5,6	24,2	10,8	79,8	64,9	71,6
Logit-Dicos	03	05	80 231	**45,2**	5,9	30,2	16,3	70,7	53,5	60,9
Logit+Test	03	05	141 550	**26,3**	3,2	18,6	8,1	86,6	73,3	79,4
Logit-D25	03	05	100 027	**36,2**	5,6	24,6	10,9	79,7	64,6	71,3
Logit-D50	03	05	73 332	**36,7**	5,4	25,2	11,0	79,5	63,8	70,8
Logit-D75	03	05	50 408	**39,0**	5,4	27,0	11,7	78,2	61,3	68,7

P : précision, **R** : rappel, **Fm** : F-mesure

Comparaisons avec d'autres systèmes

⇒ *ETAPE (provisoire), systèmes orientés connaissances et systèmes orientés données*

Part.	**Manuel**	**Rover**	**WER 23**	**WER 24**	**WER 25**	**WER 30**	**WER 35**
S1	85,6	98,1	100,7	94,2	98,9	98,4	100,9
S2	156,6	147,4	178,8	160,4	168,0	163,9	168,2
S3	36,6	**57,2**	**59,3**	64,7	**62,0**	**61,7**	**71,8**
S4	50,5	88,0	98,8	76,8	92,8	94,9	99,6
S5	44,8	69,7	73,8	72,1	73,7	74,8	86,0
S6	na	79,2	79,5	66,8	80,8	80,0	87,0
S7	na	67,8	68,4	67,6	70,9	69,9	85,2
S8	37,5	na	na	na	na	na	na
S9	62,5	75,8	79,2	76,9	79,8	80,5	90,5
S10	39,3	65,0	69,9	66,3	70,5	69,9	87,0
CasEN	**35,3**	na	na	68,4	na	na	na
mXS	38,4	63,7	67,5	**64,1**	69,1	68,6	80,4
Position	4	2	2	**1**	2	2	2

Conclusions

Conclusions

- Reconnaissance des entités nommées :
  - Problématique d'actualité devant l'afflux de données (télé, radio)
  - Expressions linguistiques **difficiles à circonscrire** (annotation)
  - Tâche qui demande des **ressources** (lexiques, transducteurs)

Conclusions

- Reconnaissance des entités nommées :
  - Problématique d'actualité devant l'afflux de données (télé, radio)
  - Expressions linguistiques **difficiles à circonscrire** (annotation)
  - Tâche qui demande des **ressources** (lexiques, transducteurs)
- Propositions et contributions :
  - $\Rightarrow$ *Décomposition de la problématique à l'aide des **marqueurs***
  - $\Rightarrow$ ***Fouille de données*** pour traiter le **langage naturel**
  - $\Rightarrow$ ***Reconnaissance séparée** du début ou de la fin d'une entité*
  - $\Rightarrow$ *Proposition de **motifs de segments***
  - $\Rightarrow$ *Cadre formel (**objectif**, **exhaustif**) pour observer le langage*
  - $\Rightarrow$ *Modèle **compétitif** et **robuste***

Conclusions

- Reconnaissance des entités nommées :
  - Problématique d'actualité devant l'afflux de données (télé, radio)
  - Expressions linguistiques **difficiles à circonscrire** (annotation)
  - Tâche qui demande des **ressources** (lexiques, transducteurs)
- Propositions et contributions :
  - ⇒ *Décomposition de la problématique à l'aide des **marqueurs***
  - ⇒ ***Fouille de données*** pour traiter le **langage naturel**
  - ⇒ ***Reconnaissance séparée** du début ou de la fin d'une entité*
  - ⇒ *Proposition de **motifs de segments***
  - ⇒ *Cadre formel (**objectif**, **exhaustif**) pour observer le langage*
  - ⇒ *Modèle **compétitif** et **robuste***
- Quelques points de la thèse non évoqués ici :
  - Expériences sur le corpus ESTER2 [LREC'10]
  - Hybridation [Workshop EACL'12]

Perspectives

Perspectives

- Amélioration de l'approche :
  - **Paramétres optimaux** pour le système mXS (Logit)
  - Modèles envisageables et performants au niveau **local**
  - **Vraisemblance globale** (interactions EN, anaphores, catégorisation verbale)

Perspectives

- Amélioration de l'approche :
  - **Paramétres optimaux** pour le système mXS (Logit)
  - Modèles envisageables et performants au niveau **local**
  - **Vraisemblance globale** (interactions EN, anaphores, catégorisation verbale)
- Cadre de l'approche :
  - Données **séquentielles**
  - Données potentiellement **riches** (ontologies)
  - Recherche de **structuration** (schéma : annotation ou autre)
  - Disponibilité de données

Quelques pistes...

Quelques pistes...

$\Rightarrow$ *Amélioration des estimations locales*

- Souhaite-t-on des probabilités ou un classement ?
- Peut-on mieux tenir compte du poids des erreurs commises ?
- Quels marqueurs d'entités nommées se substituent le mieux ?

Quelques pistes...

⇒ *Amélioration des estimations locales*

- Souhaite-t-on des probabilités ou un classement ?
- Peut-on mieux tenir compte du poids des erreurs commises ?
- Quels marqueurs d'entités nommées se substituent le mieux ?

⇒ *Expérimenter les motifs de segments*

- Préciser les conditions qui permettent un gain en performances ?
- Limitation de l'explosion combinatoire des séquences ?
- Liens avec les grammaires probabilistes (constituants) ?

Quelques pistes...

⇒ *Amélioration des estimations locales*

- Souhaite-t-on des probabilités ou un classement ?
- Peut-on mieux tenir compte du poids des erreurs commises ?
- Quels marqueurs d'entités nommées se substituent le mieux ?

⇒ *Expérimenter les motifs de segments*

- Préciser les conditions qui permettent un gain en performances ?
- Limitation de l'explosion combinatoire des séquences ?
- Liens avec les grammaires probabilistes (constituants) ?

- Ouverture à d'autres problématiques :
  - Séquences ADN
  - Logs de sessions
  - Structuration web
  - ...

Merci pour votre attention !

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, and Denis Maurel.

An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign.
In *International Language Resources and Evaluation (LREC'10)*, 2010.

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, and Arnaud Soulet.

Recognizing named entities using automatically extracted transduction rules.
In *Language & Technology Conference (LTC'11)*, 2011.

Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger, and Arnaud Soulet.

Coupling knowledge-based and data-driven systems for named entity recognition.
In *Innovative hybrid approaches to the processing of textual data (EACL'12 workshop)*, 2012.

Damien Nouvel, jean Yves Antoine, Denis Maurel, and Nathalie Friburger.

Reconnaissance d'entités nommées, nouvelles frontières et nouvelles approches.
Journée de l'ATALA - Reconnaissance d'Entités Nommées, Nouvelles Frontières et Nouvelles Approches, 2011.

Damien Nouvel and Arnaud Soulet.

Annotation d'entités nommées par extraction de règles de transduction.
In *Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'10)*, 2011.

Damien Nouvel, Arnaud Soulet, Jean-Yves Antoine, Nathalie Friburger, and Denis Maurel.

Reconnaissance d'entités nommées : enrichissement d'un système à base de connaissances à partir de techniques de fouille de textes.
In *Traitement Automatique des Langues Naturelles (TALN'10)*, 2010.

Compléments

- Adaptations TreeTagger [Schmid 1994] :
  - **Verbes** : suppression du mode et du temps.
  - **Noms propres et abréviations** : catégorie `'NAMABR'`
  - **Déterminants** : amalgames `'PRPDET'`, déterminants `'DET/DEF'`
  - **Nombres** : sous-catégorisés selon leurs chiffres `'NUM/DIGITS:1'` ...`'NUM/DIGITS:4'` (préfixes années) `'NUM/DIGITS:MANY'`
  - **NAMABR, noms, verbes** : catégorie par suffixe (trois caractères)

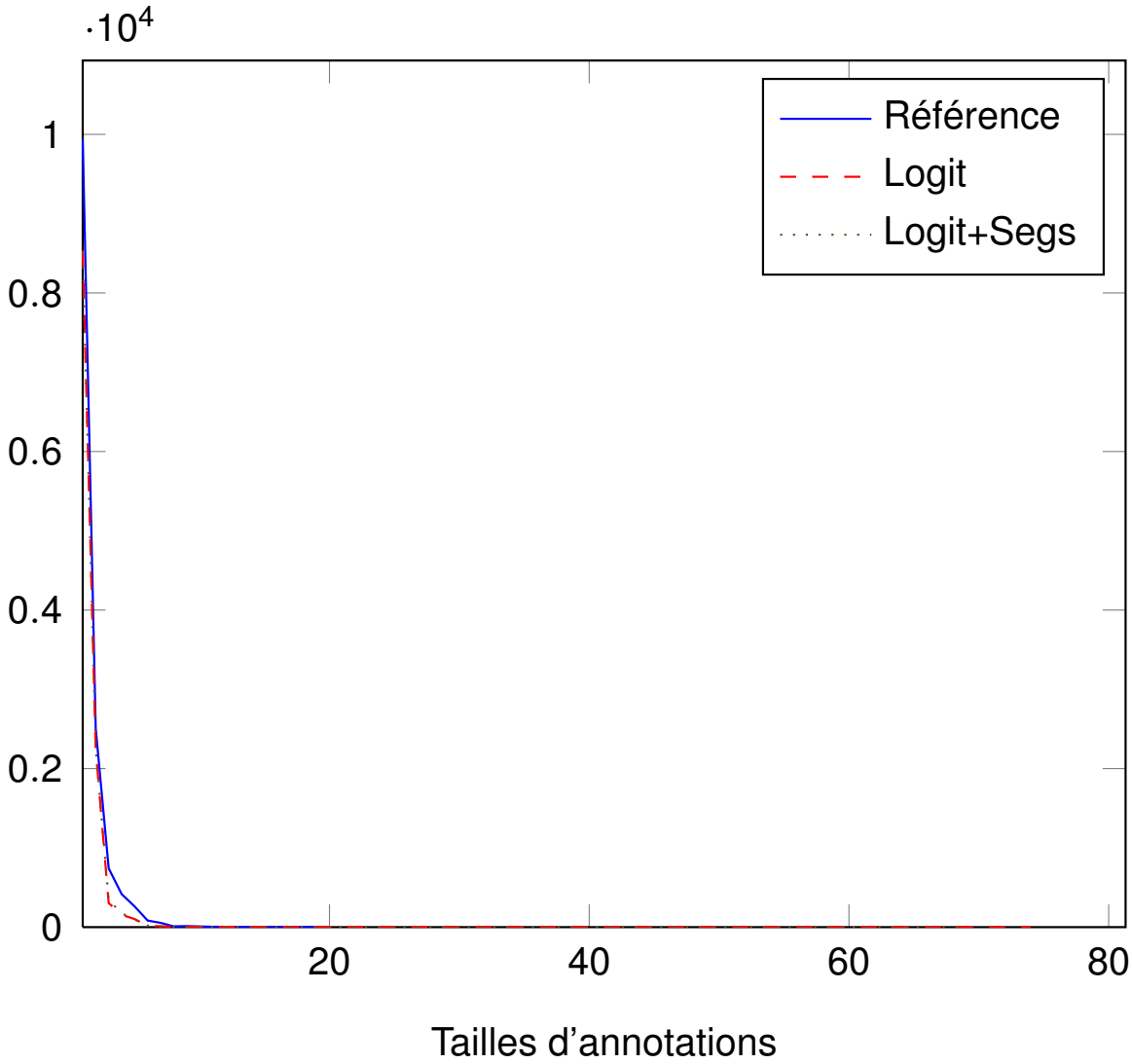
$\Rightarrow$ *Peu d'impact sur la théorie, beaucoup plus sur les performances*

$\Rightarrow$ *Reconnaissance en prétraitement des* ***noms propres***

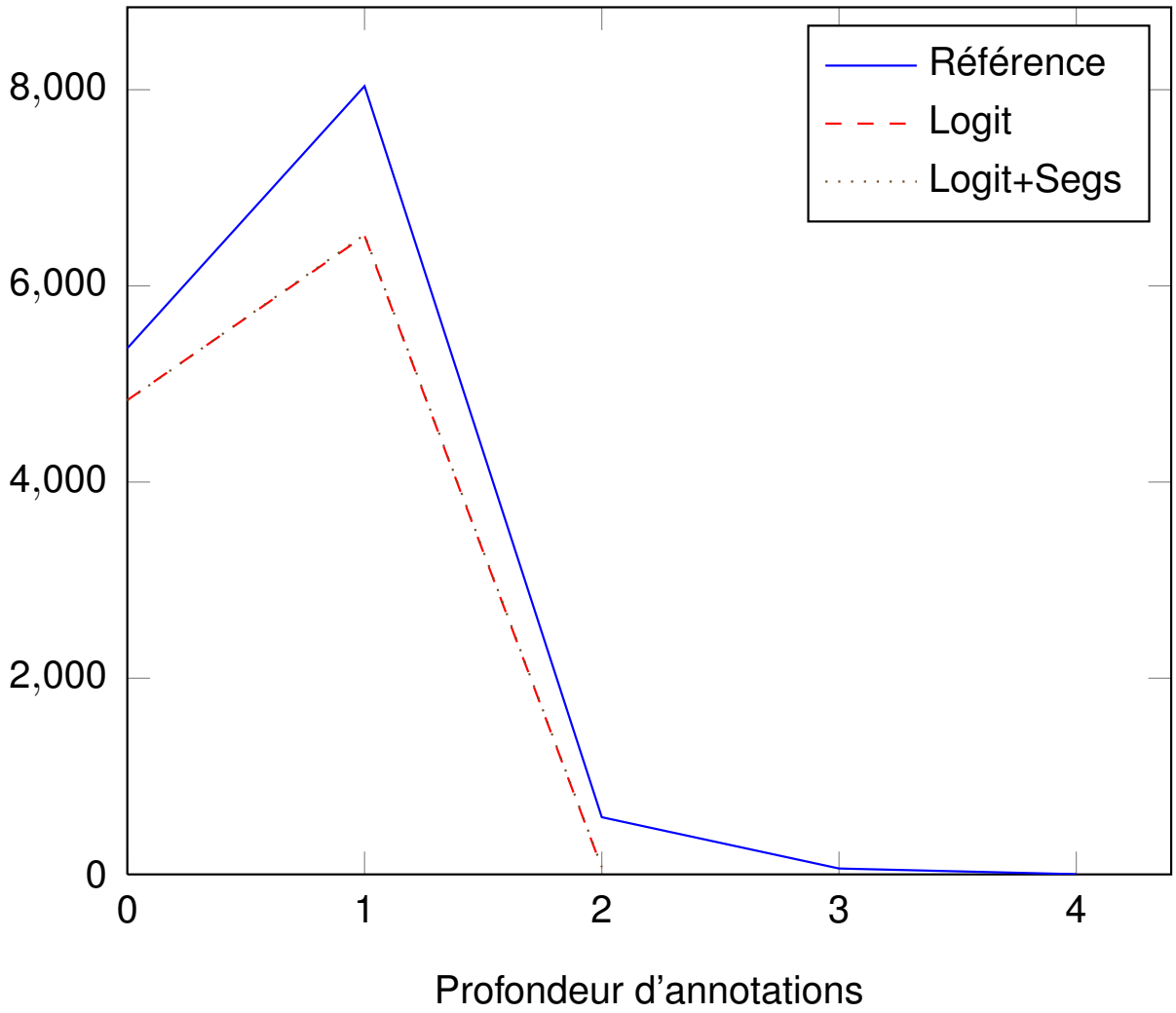
Exemple de prétraitement

```
'PRO/PER/je/Je VER/pres/suivre/suis⊕VER/pres/être/suis
PRPDET/au/au NAMABR/NP/Centre NAMABR/NP/Georges/Georges
NAMABR/NP/Pompidou/Pompidou SENT/./.'
```

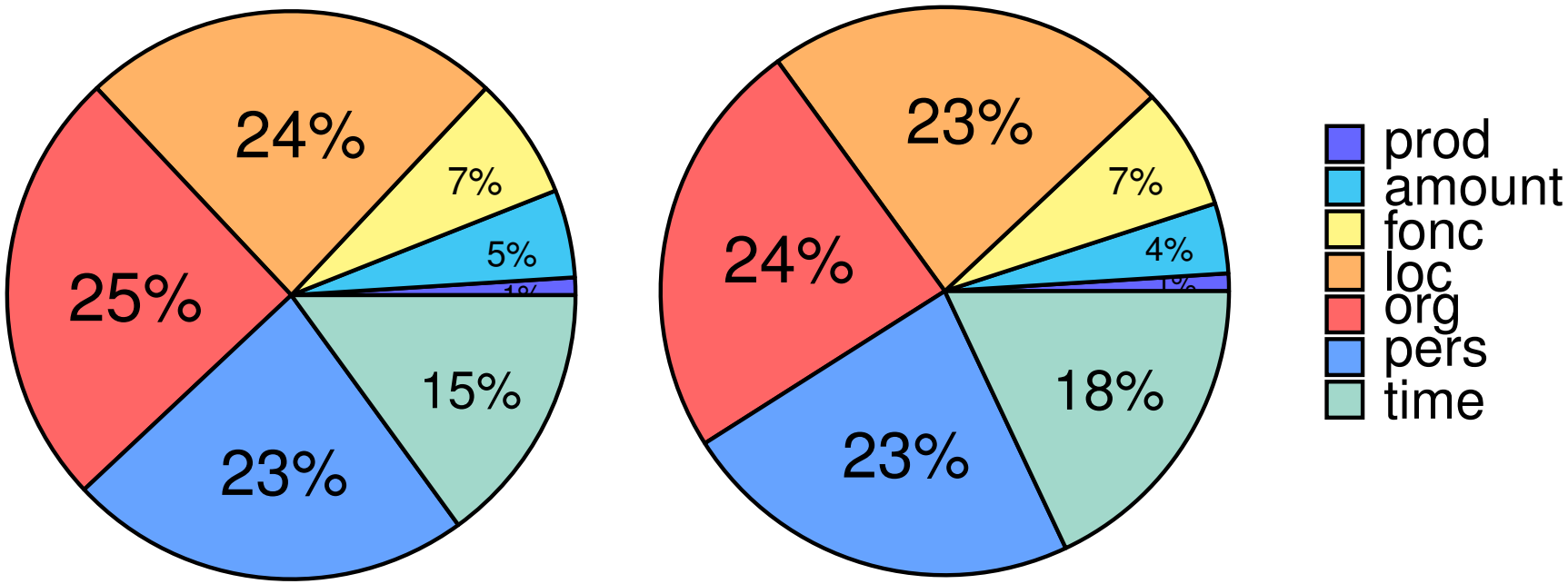
Compléments



Compléments



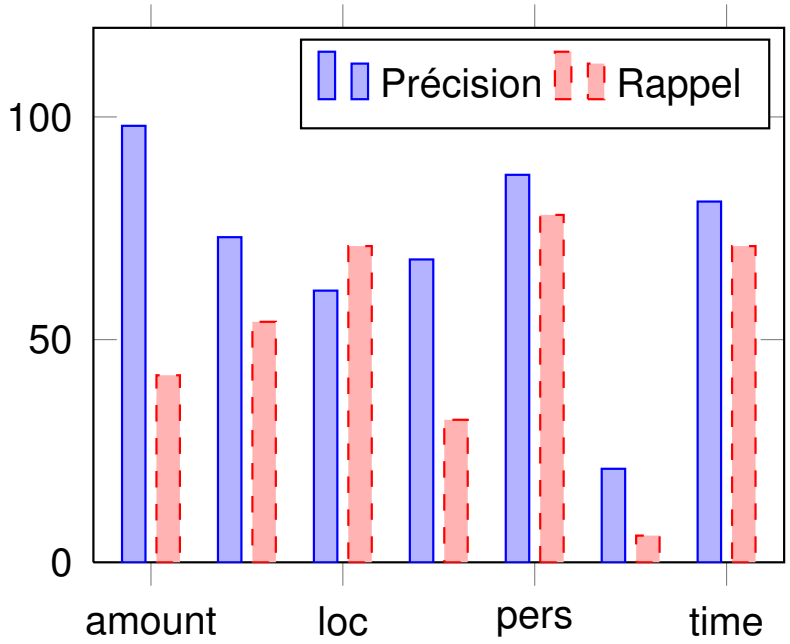
Compléments



(k) Ester2-Dev (l) Ester2-Test

FIGURE : Répartition des types d'entités nommées pour chaque partie d'Ester2

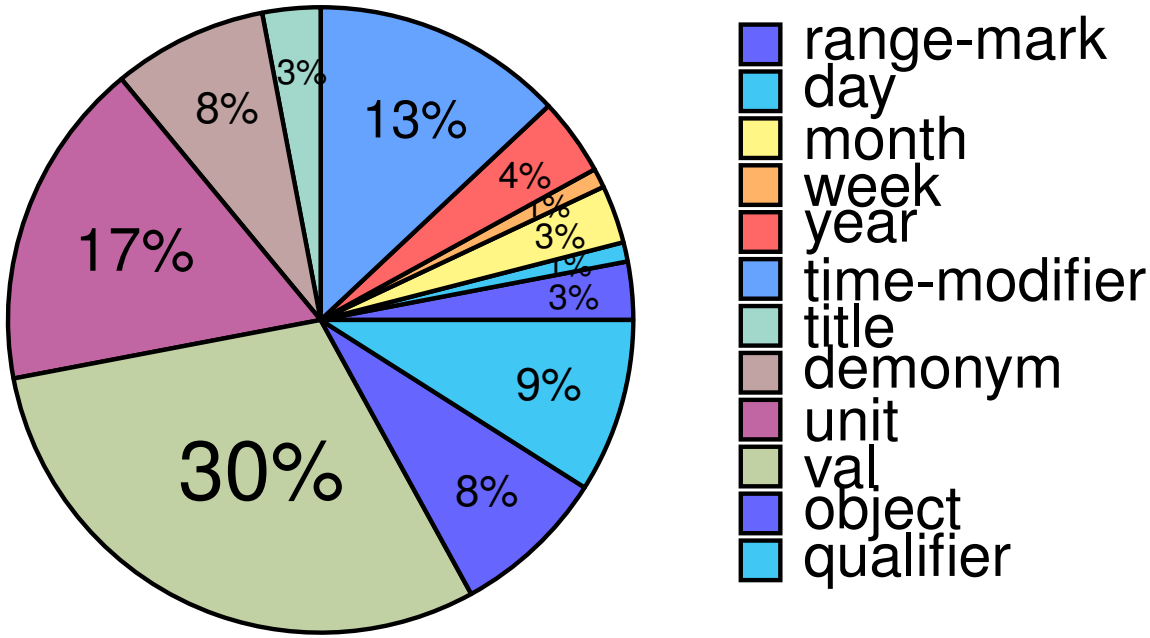
Compléments



(a) Précision et rappel

FIGURE : Performance de `CasEN` par type d'entités nommées lors de la campagne Ester2

Compléments



(a) Etape-Test

FIGURE : Répartition des types de composants pour chaque partie d'Etape

Compléments

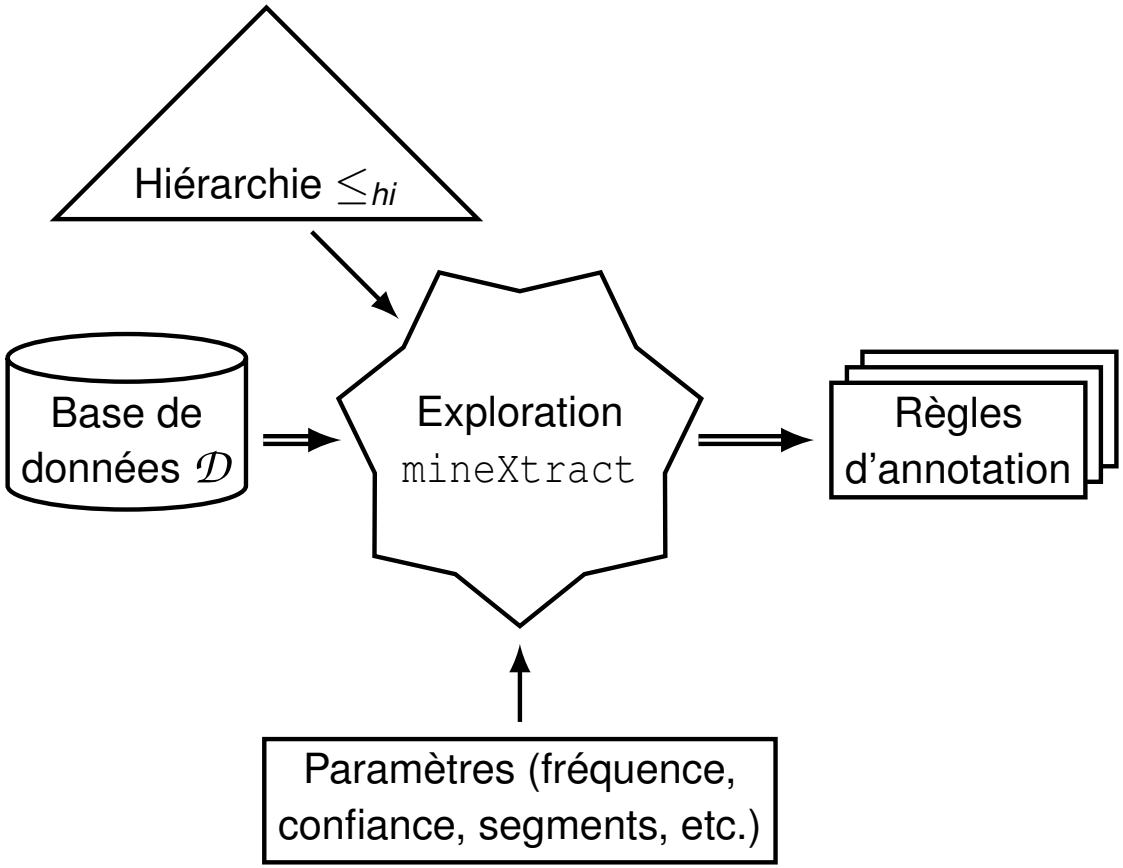


FIGURE : Processus d'extraction de motifs

Compléments

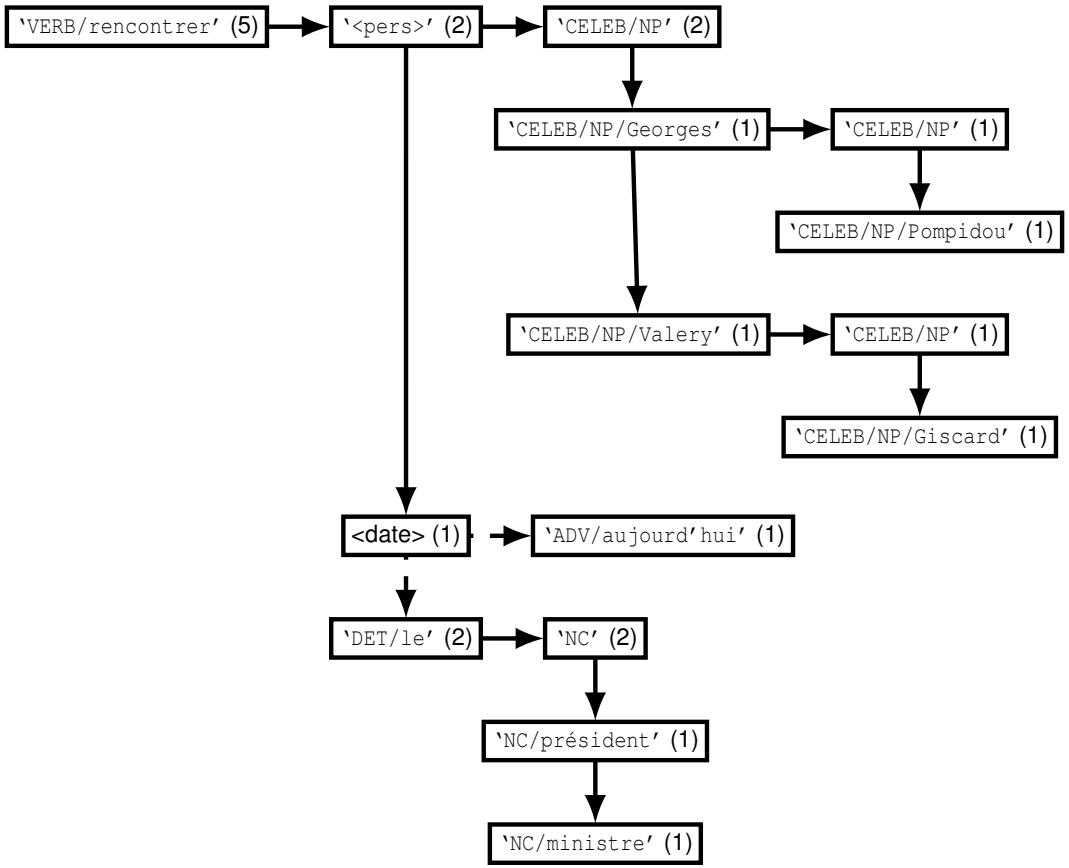
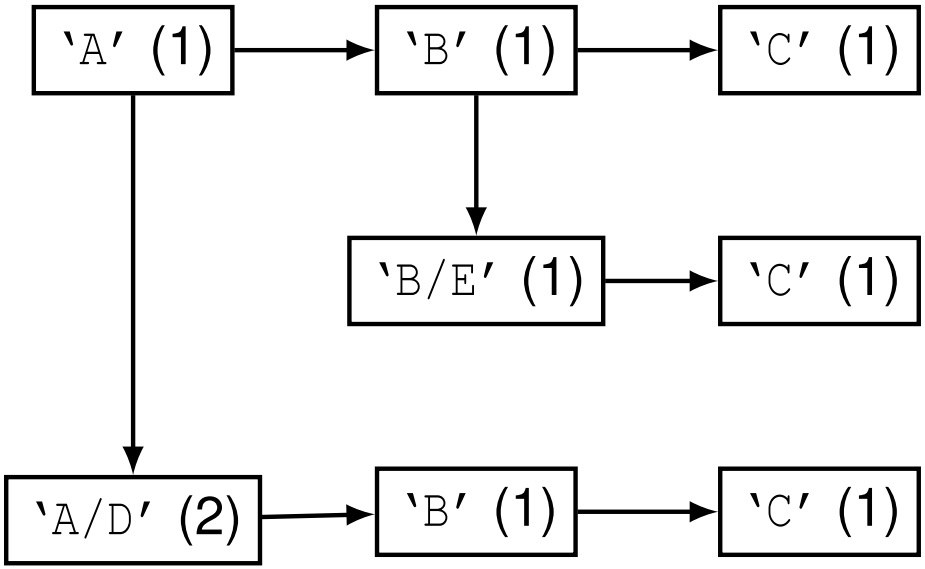
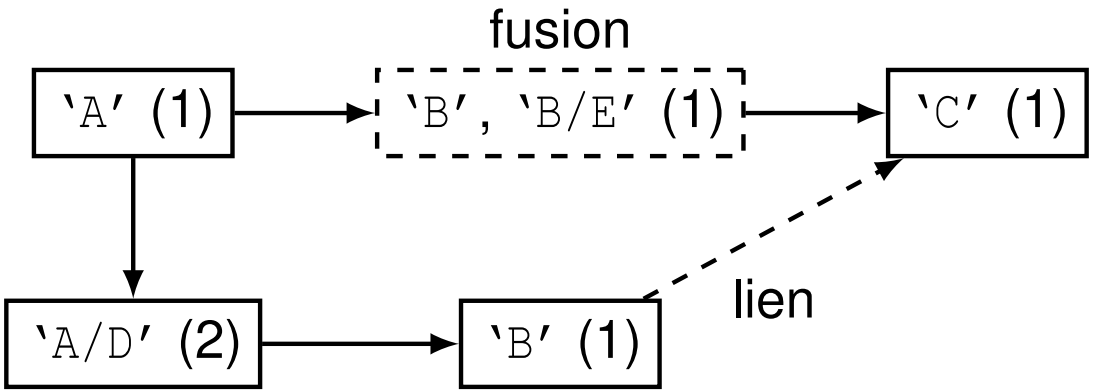


FIGURE : Arbre des préfixes communs

Compléments

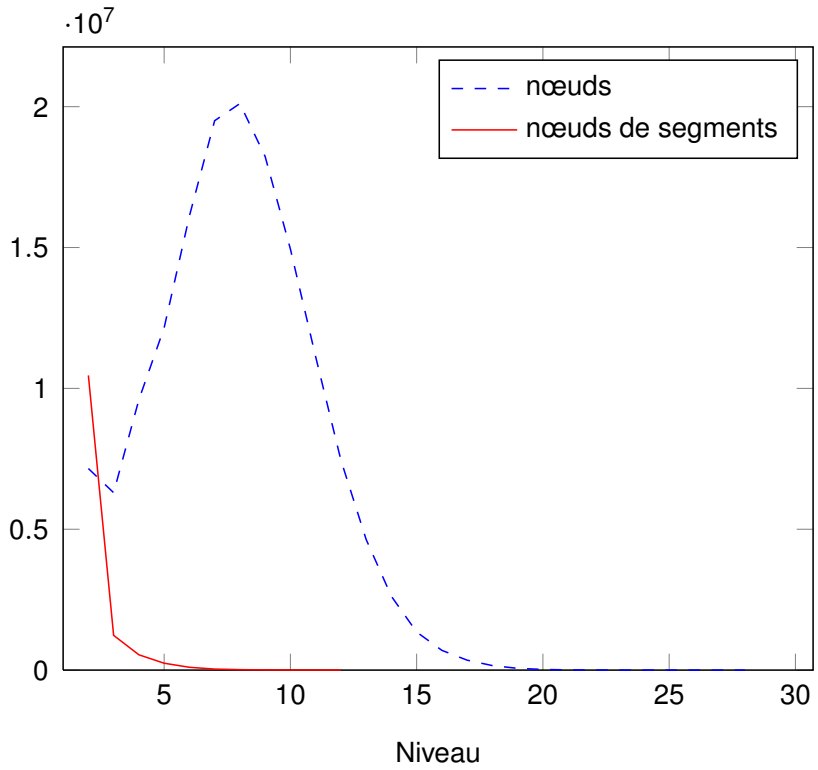


(a) Avant optimisations



(b) Après optimisations

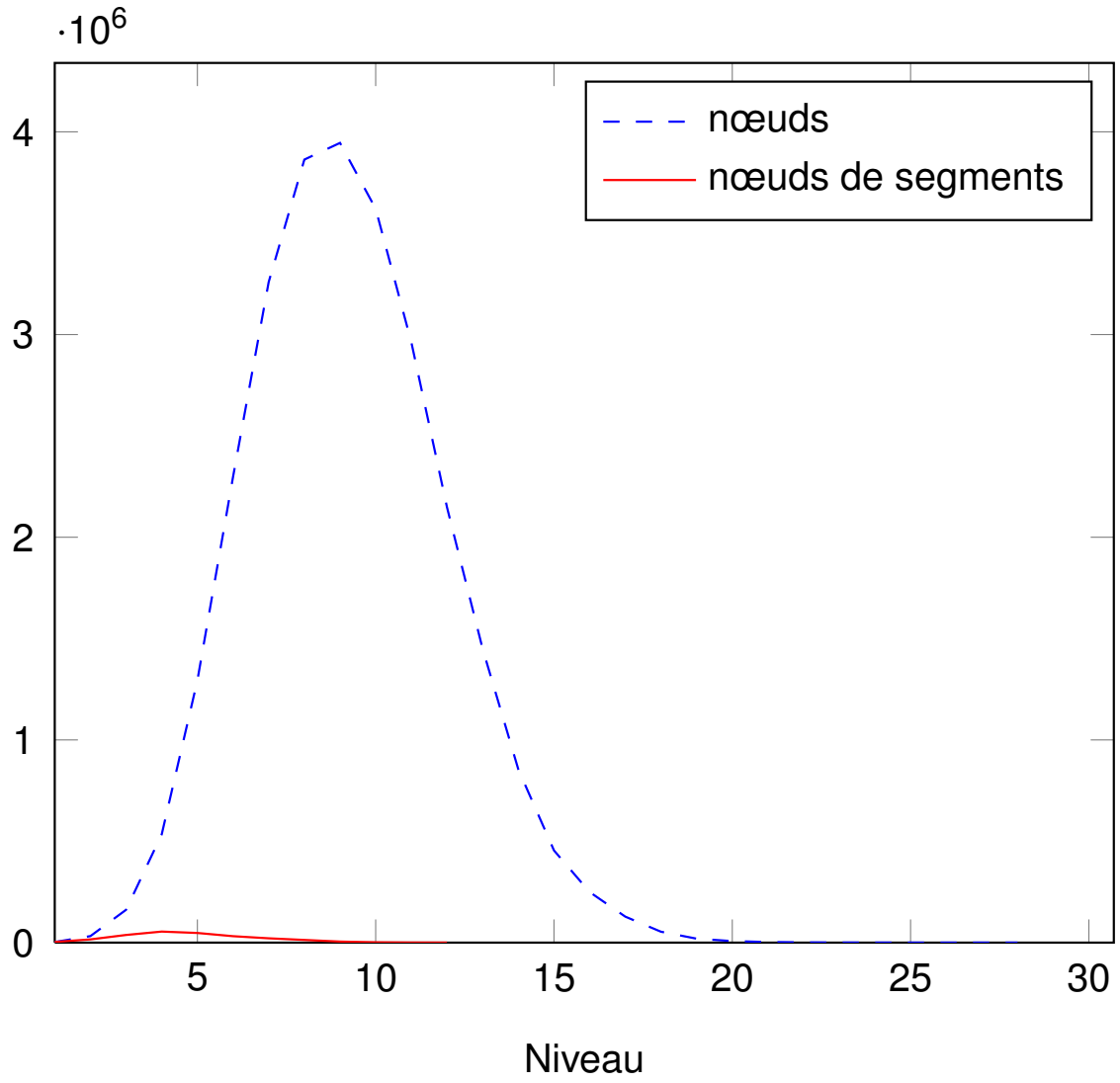
Compléments



(a) Candidats

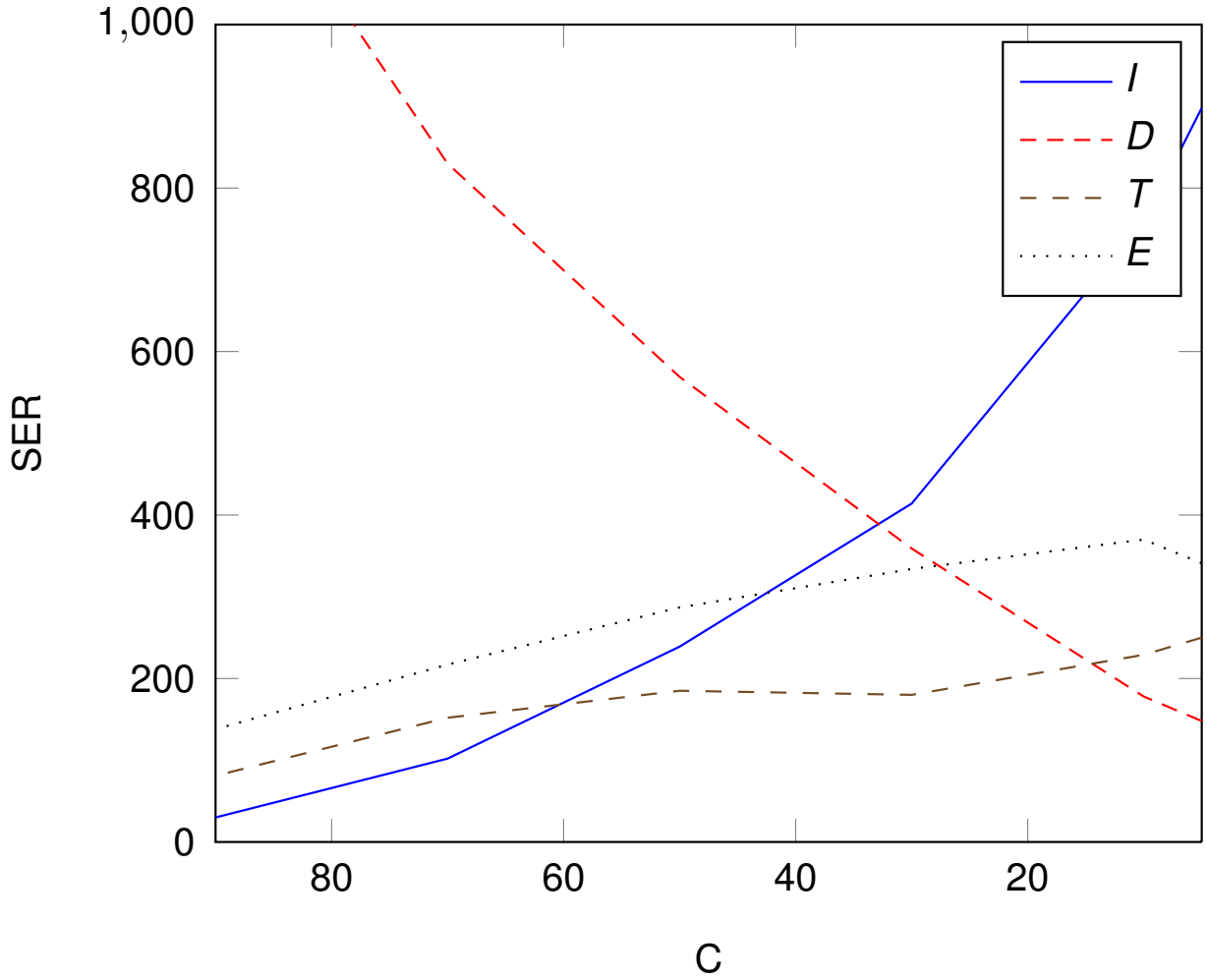
FIGURE : Nombre de nœuds (849 385 règles) ou de nœuds de segments (15 103 règles) sur Ester2 à fréquence $F \geq 12$ et confiance $C \geq 0,1$ après enrichissment syntaxique

Compléments



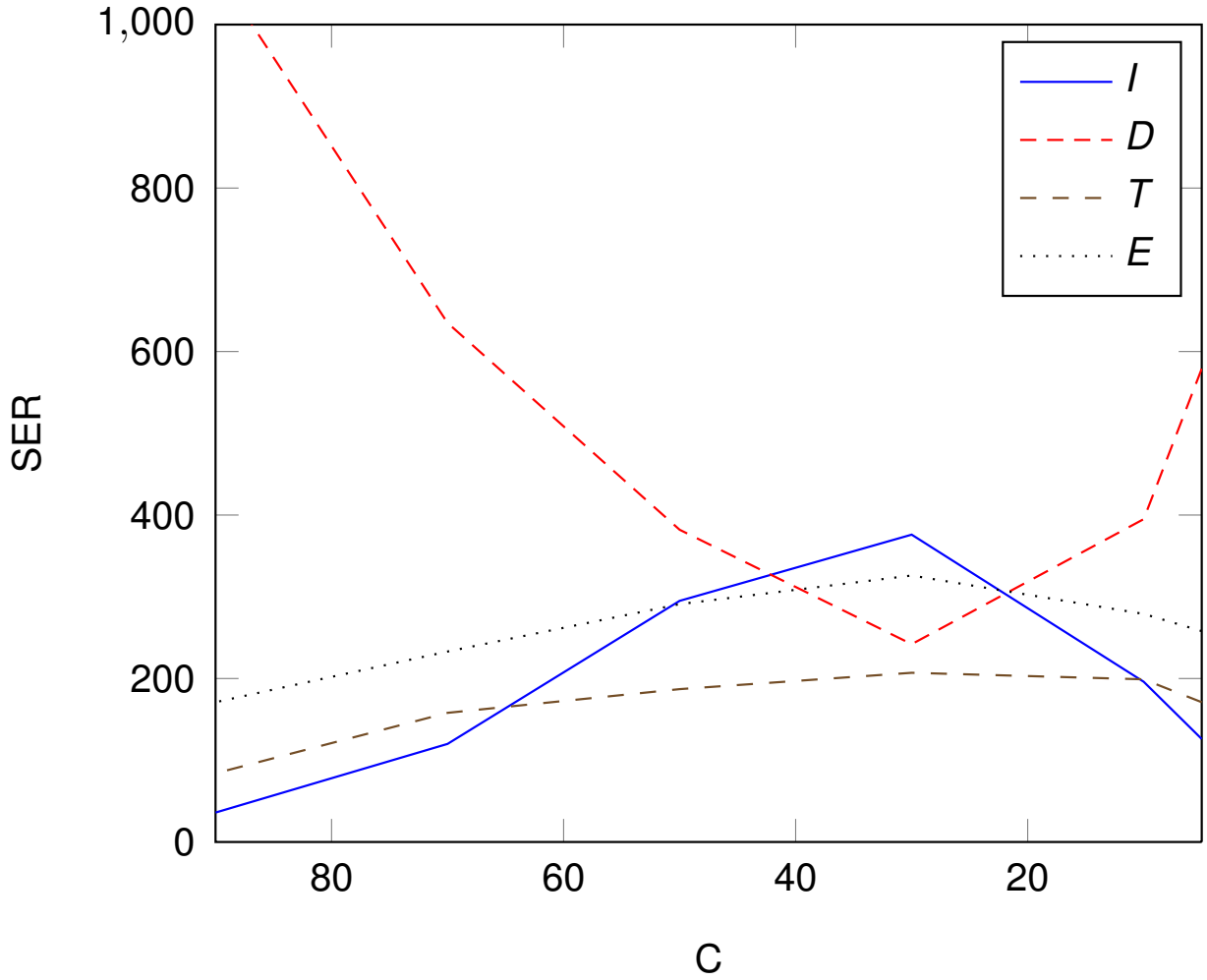
(a) Fréquents

Compléments



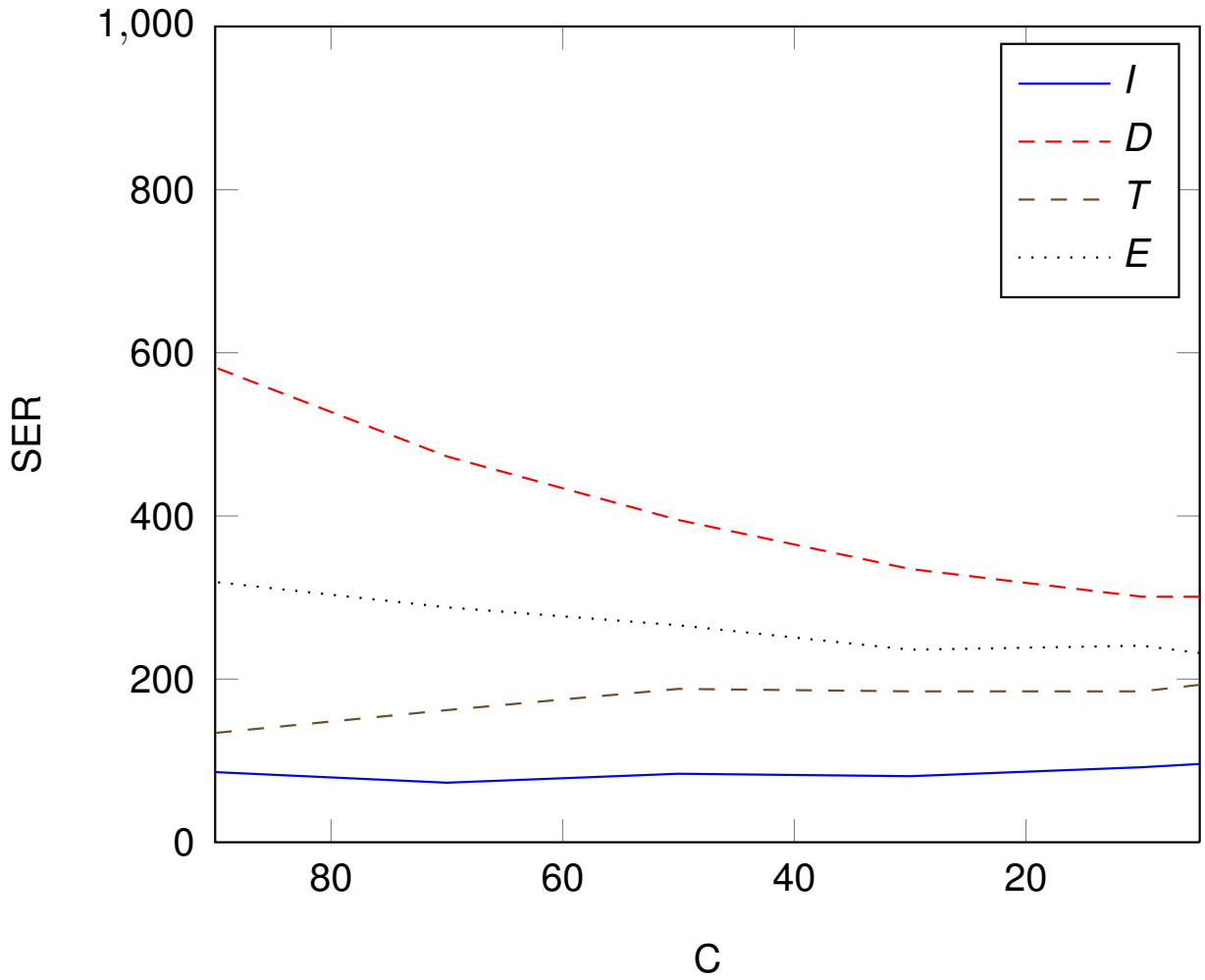
(a) Erreurs, $F \geq 5.10^{-5}$

Compléments



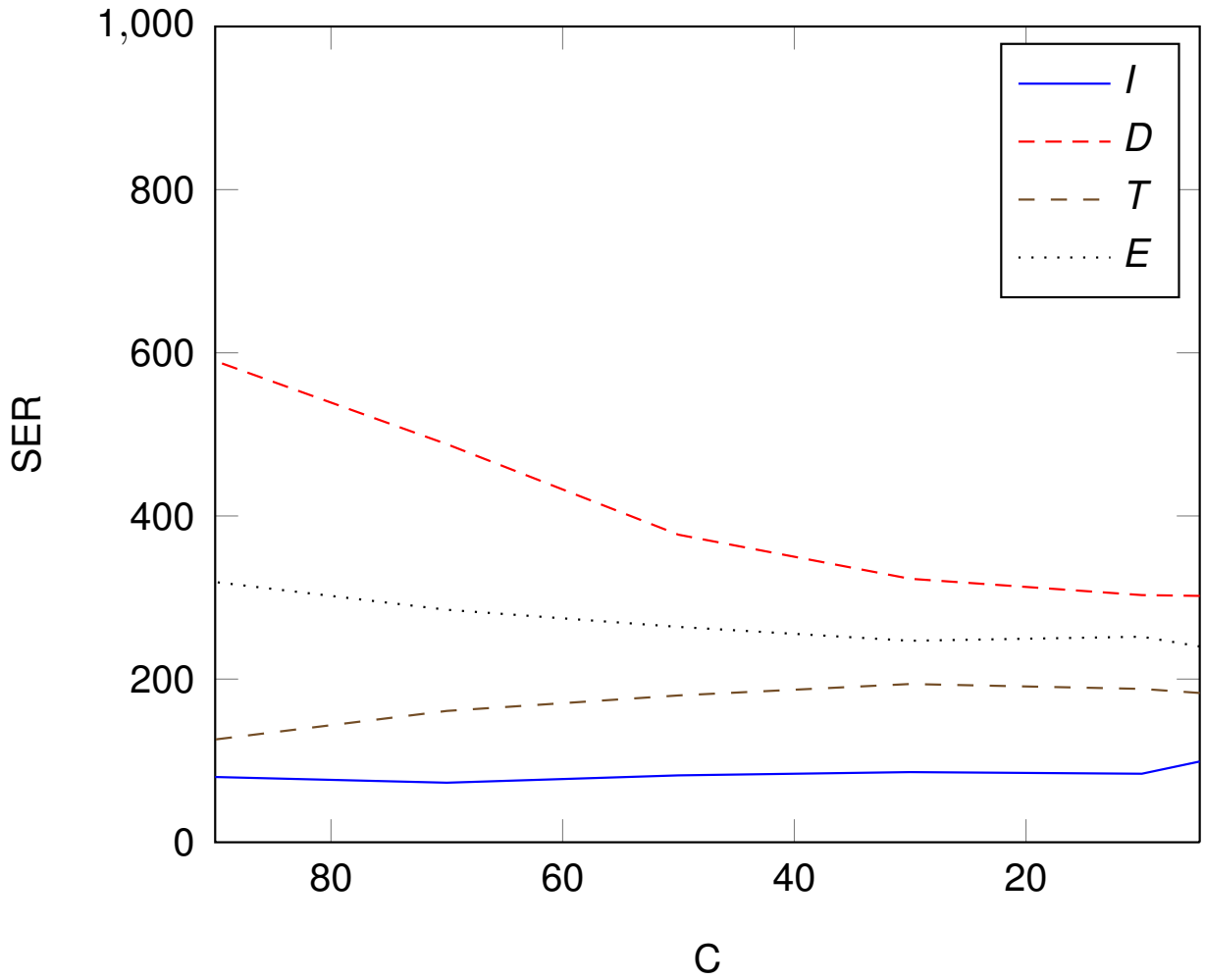
(a) Erreurs, $F \geq 5.10^{-5}$

Compléments



(a) Erreurs, $F \geq 5.10^{-5}$

Compléments



(a) Erreurs, $F \geq 5.10^{-5}$

Compléments

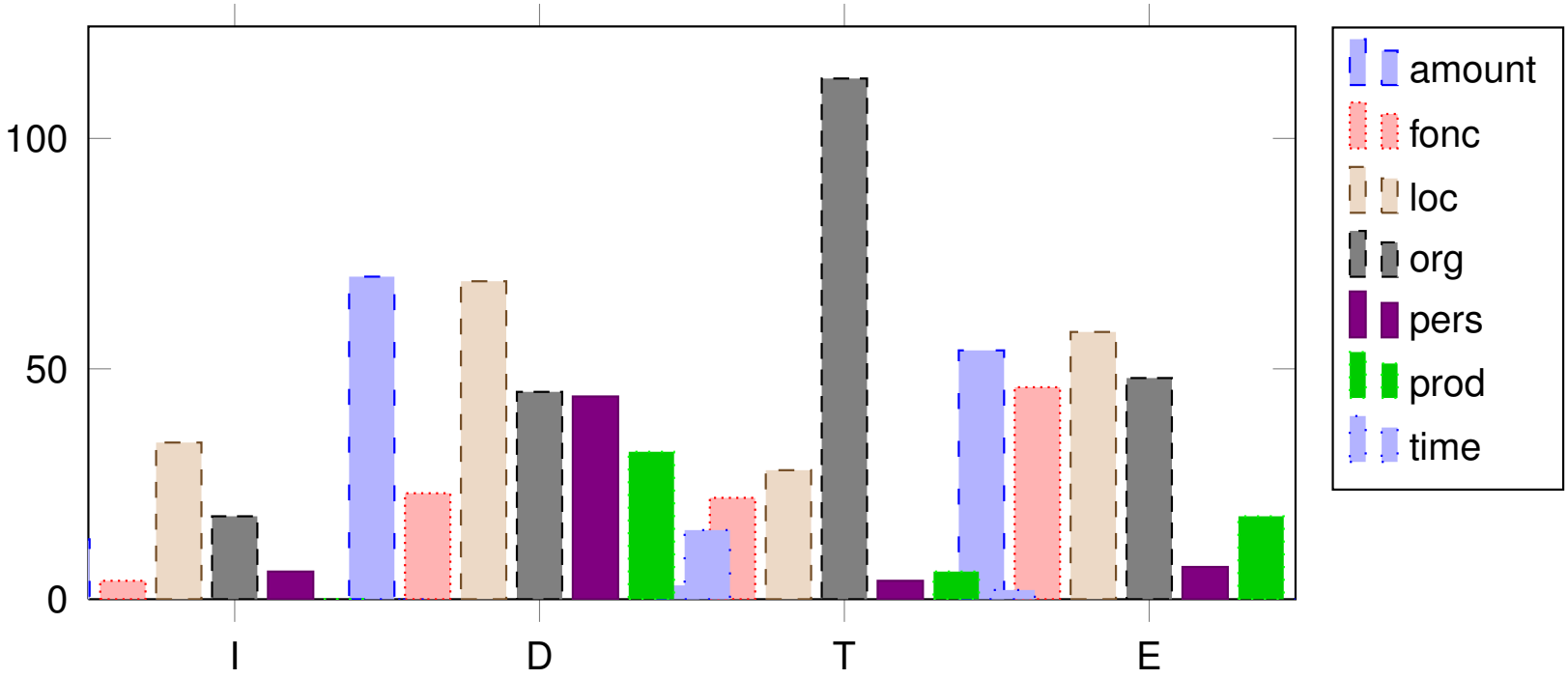
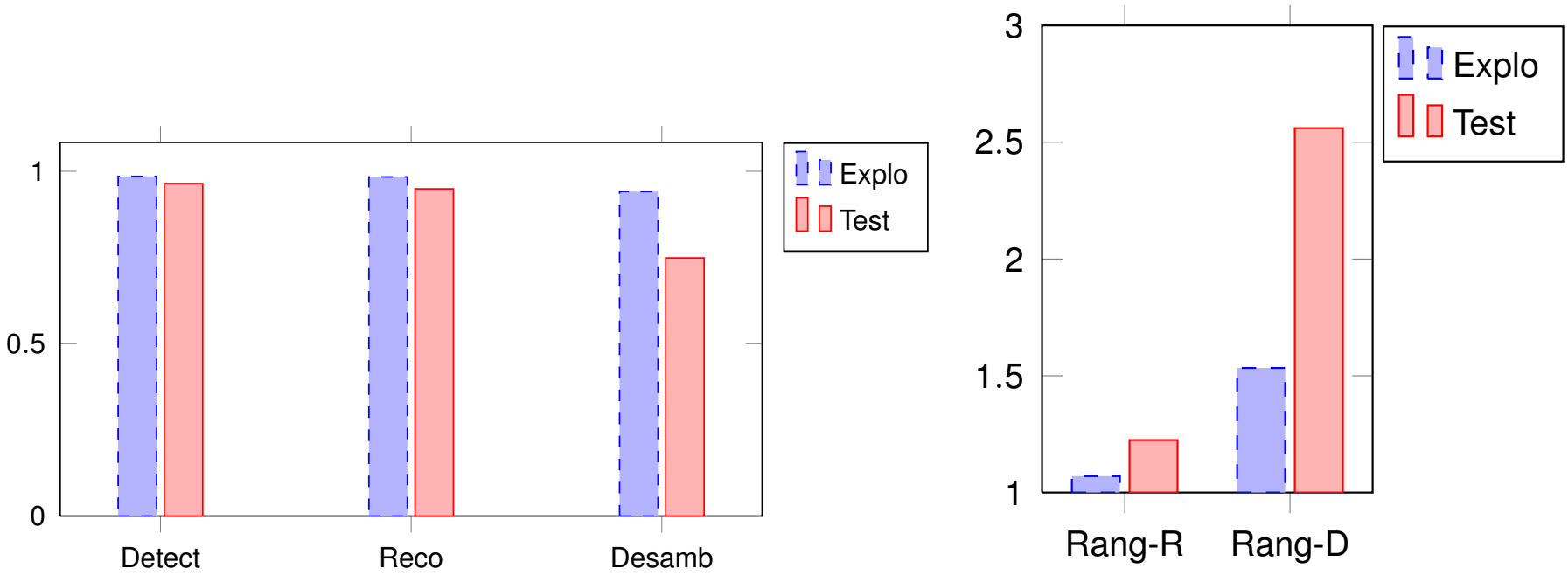


FIGURE : Erreurs par type sur Ester2

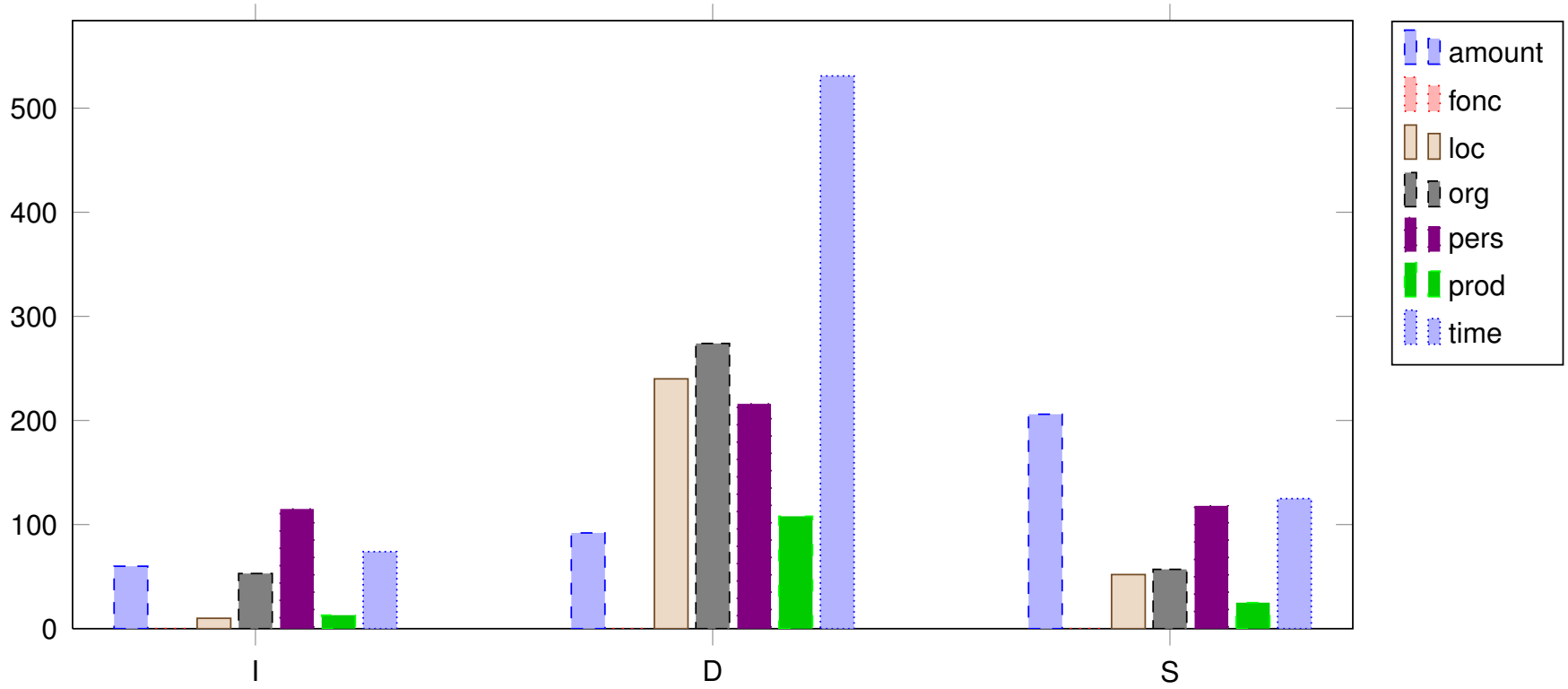
Compléments



(a) Taux de détection, reconnaissance et désambi- (b) Rang moyen des séquences guisation

FIGURE : Détection, reconnaissance, désambiguisation et ordonnancement sur le corpus exploré (Explo) et sur le corpus de test (Test) avec l'approche `Logit` sur Ester2

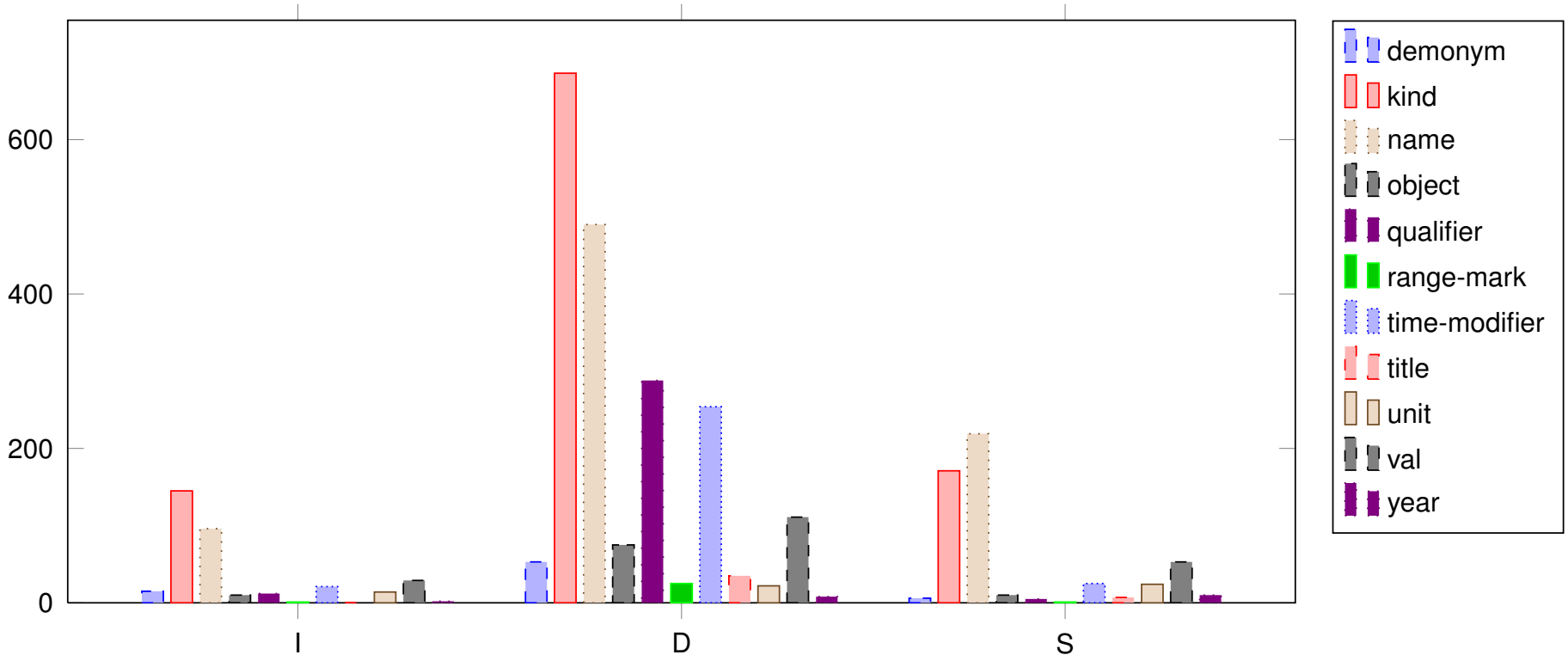
Compléments



(a) Entités

FIGURE : Erreurs par types primaires d'entités nommées et de composants sur Ester2

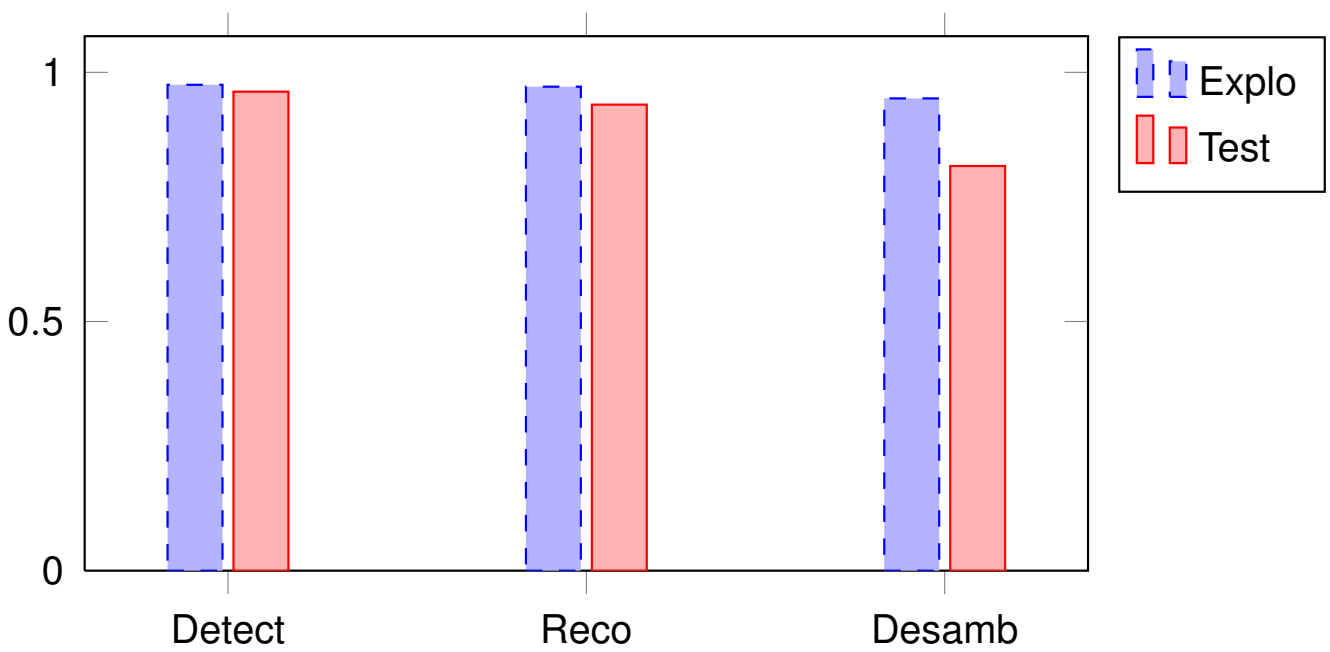
Compléments



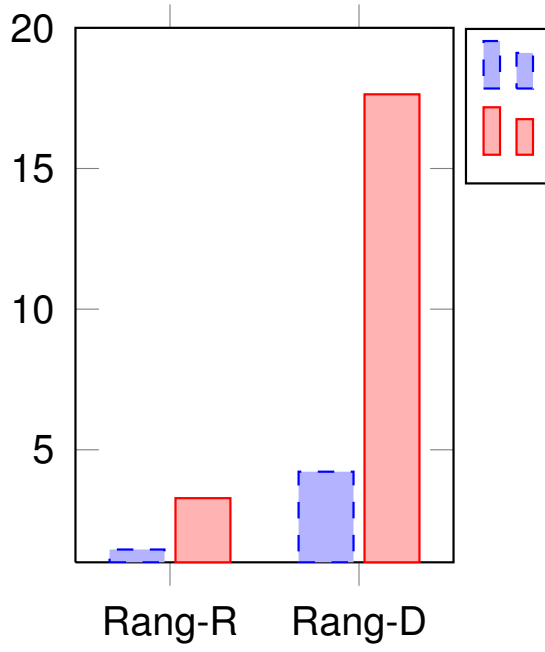
(a) Composants

FIGURE : Erreurs par types primaires d'entités nommées et de composants sur Ester2

Compléments



(a) Taux de détection, reconnaissance et désambiguisation



(b) Rang moyen des séqu

FIGURE : Détection, reconnaissance, désambiguisation et ordonnancement sur le corpus exploré (Explo) et sur le corpus de test (Test) avec l'approche `Logit` sur Etape