Désambiguïsations lexicales
Méthodes et résultats

Damien Nouvel

Plan

1. La désambiguisation lexicale : des formes aux sens

2. Reconnaissance et liaison des entités nommées

3. Enrichissement lexical d'un corpus bambara

4. Quelques travaux préliminaires sur l'arabe

5. Conclusions

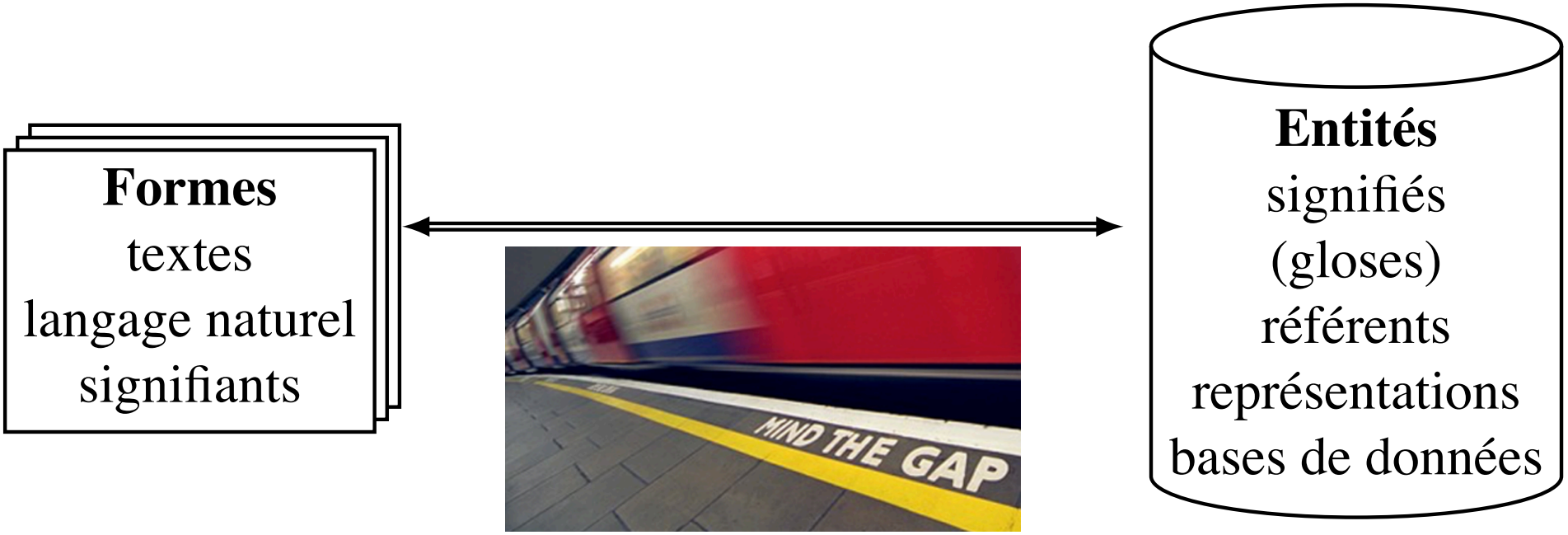
Des formes aux sens ...

Des formes aux sens ...

Formes
textes
langage naturel
signifiants

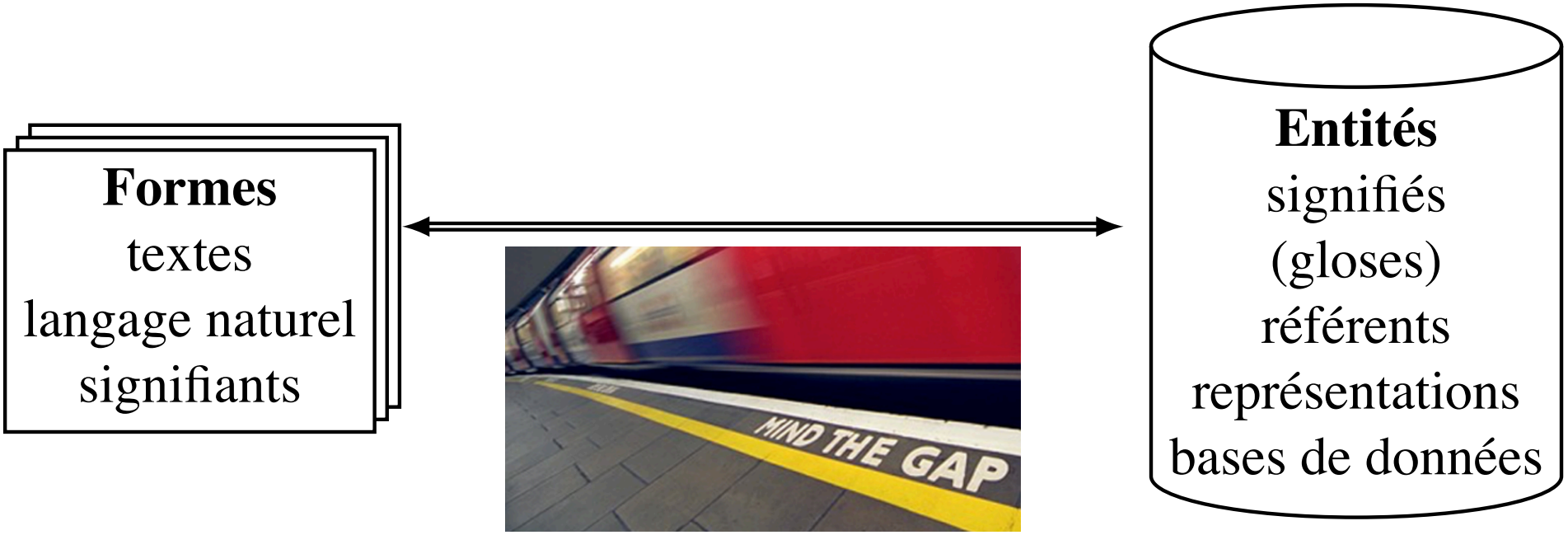
Entités
signifiés
(gloses)
référents
représentations
bases de données

Des formes aux sens ...



- Comment faire le **lien** entre **textes** et **bases de connaissances** ?
- Quel **niveau** d'analyse : caractères, mots, énoncés, documents ?
- Comment représenter le **contexte** ?

Des formes aux sens ...



- Comment faire le **lien** entre **textes** et **bases de connaissances** ?
- Quel **niveau** d'analyse : caractères, mots, énoncés, documents ?
- Comment représenter le **contexte** ?

⇒ Processus de **compréhension** (vs génération)

On ne peut toujours comprendre

« Il a regardait la **pioche** posée juste devant lui …

On ne peut toujours comprendre

« Il a regardait la **pioche** posée juste devant lui …
l'enjeu était important, et il essayait d'évaluer le risque …

On ne peut toujours comprendre

« Il a regardait la **pioche** posée juste devant lui …
l'enjeu était important, et il essayait d'évaluer le risque …
il s'agissait de ne pas se tromper, cette fois-ci …

On ne peut toujours comprendre

« Il a regardait la **pioche** posée juste devant lui …
l'enjeu était important, et il essayait d'évaluer le risque …
il s'agissait de ne pas se tromper, cette fois-ci …
les autres étaient assemblés autour de lui, en silence …

On ne peut toujours comprendre

« Il a regardait la **pioche** posée juste devant lui …
l'enjeu était important, et il essayait d'évaluer le risque …
il s'agissait de ne pas se tromper, cette fois-ci …
les autres étaient assemblés autour de lui, en silence …
puis, finalement, il se décida à agir ! »

⇒ On ne peut toujours comprendre avec 100% de certitude …

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  $\Rightarrow$ Quelle existence pour le langage sans communication ?

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  ⇒ Quelle existence pour le langage sans communication ?
- Communication interpersonnelle (« proches »)
  - Dialogue en interaction : peu d'ambiguïtés

  ⇒ Aspect visuel (gestes, visage, etc.)

  ⇒ Contextualisation des énoncés

  ⇒ À l'écrit : des langages **prescrits** vs **décrits**

Désambiguisation : situations et enjeux

- La désambiguisation comme **compréhension**
- Sans communication
  - Sens et représentation dans le **cerveau** (imagerie)
  - Représentations cognitives / **neuronales**

  ⇒ Quelle existence pour le langage sans communication ?
- Communication interpersonnelle (« proches »)
  - Dialogue en interaction : peu d'ambiguïtés

  ⇒ Aspect visuel (gestes, visage, etc.)

  ⇒ Contextualisation des énoncés

  ⇒ À l'écrit : des langages **prescrits** vs **décrits**
- Variations de l'intercommunication
  - Prononciations, dialectes, bruit
  - Le langage comme **convention sociale**
  - Langues de **spécialité** (domaines)
  - Grande diversité des **systèmes d'écritures** (150)

Plan

1. La désambiguisation lexicale : des formes aux sens

2. Reconnaissance et liaison des entités nommées
   - Les entités nommées et leur annotation
   - Reconnaître automatiquement les entités nommées
   - Règles d'annotation et exploration de données
   - Utilisation des règles d'annotation
   - Vers la résolution

3. Enrichissement lexical d'un corpus bambara

4. Quelques travaux préliminaires sur l'arabe

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

- 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)

Les entités nommées pour ... tous !

Les entités nommées pour ... tous !

Google

quel âge a Maître Gims ?

Connexion

Tous | Actualités | Vidéos | Images | Shopping | Plus | Paramètres | Outils

Environ 124 000 résultats (0,52 secondes)

Maître Gims / Âge

30 ans

6 mai 1986

Black M
31 ans

Soprano
37 ans

Kendji Girac
20 ans

Commentaires

Maître Gims — Wikipédia
https://fr.wikipedia.org/wiki/**Maître_Gims**
Maître Gims, de son vrai nom Gandhi Djuna, est un rappeur, chanteur et compositeur congolais né le **6 mai 1986** à Kinshasa. Il est issu d'une famille de ...

Maître Gims

Rappeur

Disponible sur

- YouTube
- Spotify
- Deezer
- TuneIn

Maître Gims, de son vrai nom Gandhi Djuna, est un rappeur, chanteur et compositeur congolais né le 6 mai 1986 à Kinshasa. Wikipédia

Naissance : 6 mai 1986 (30 ans), Kinshasa, République démocratique du Congo

Difficultés (Washington)

- Phénomènes liés à la langue naturelle

Difficultés (Washington)

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

Exemples ambigus

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

Difficultés (Washington)

- 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.'* (*pers*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*pers*)
*'Nous avons visité **Washington** et Boston.'* (*ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*vehic.*)

Difficultés (Washington)

- 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.'* (*pers*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*pers*)
*'Nous avons visité **Washington** et Boston.'* (*ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*vehic.*)
*'Hier, **Washington** a battu New York 34 à 10.'* (*équipe*)
*'Lors des discussions, **Washington** a opposé son veto.'* (*gouv.*)

Difficultés (Washington)

- 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.'* (*pers*)
*'Le **1er président des Etats-Unis** a rédigé la constitution.'* (*pers*)
*'Nous avons visité **Washington** et Boston.'* (*ville*)
*'Le **Washington** inaugure la ligne Le Havre - New York.'* (*vehic.*)
*'Hier, **Washington** a battu New York 34 à 10.'* (*équipe*)
*'Lors des discussions, **Washington** a opposé son veto.'* (*gouv.*)

$\Rightarrow$ Comment traiter automatiquement ce cas ?

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

$\Rightarrow$ Définitions applicatives ou en extension

Que sont les entités nommées ?

- Aux origines : campagnes MUC (DARPA, 1995-2000)
- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Désignateurs rigides d'entités » (Kripke)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - …

⇒ Définitions applicatives ou en extension

- Proposition : objets **à la frontière de la logique**
  « Les entités nommées, lorsqu'elles sont résolues, semblent désigner des objets mentaux de manière stable, à partir desquels il est attendu qu'une représentation logique opère »

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)
  - **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.

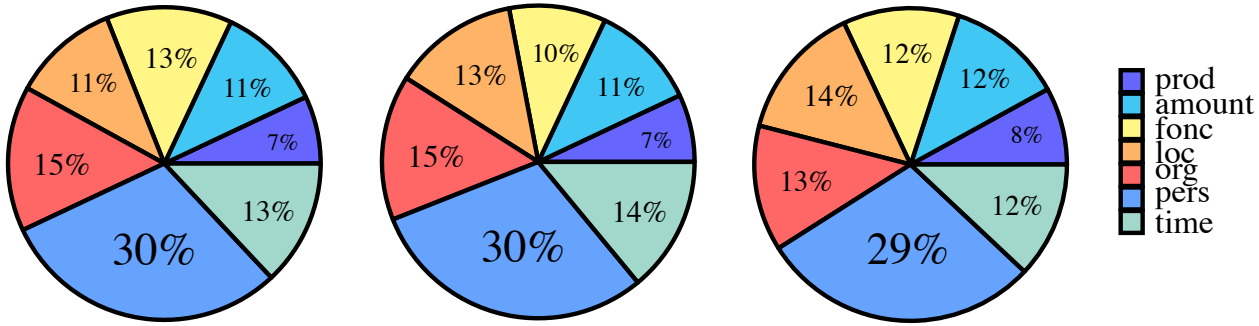
⇒ Transcriptions manuelles et automatiques (WER>20)

⇒ Annotation manuelle, projection et évaluation

Corpus ETAPE

- 74 enregistrements annotés (BFMTV, France Inter, LCP, TV8)

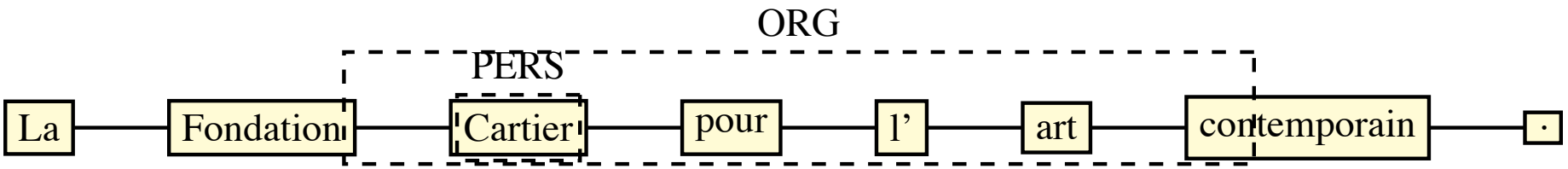
Corpus	Tokens	Enoncés	EN
Etape-Train	355 975	14 989	46 259
Etape-Dev	115 530	5 724	14 112
Etape-Test	123 221	6 770	13 055
Total	594 726	27 483	73 426



(a) Etape-Train (b) Etape-Dev (c) Etape-Test

$\Rightarrow$ 3 entités par énoncé / 1 entité tous les 7 mots

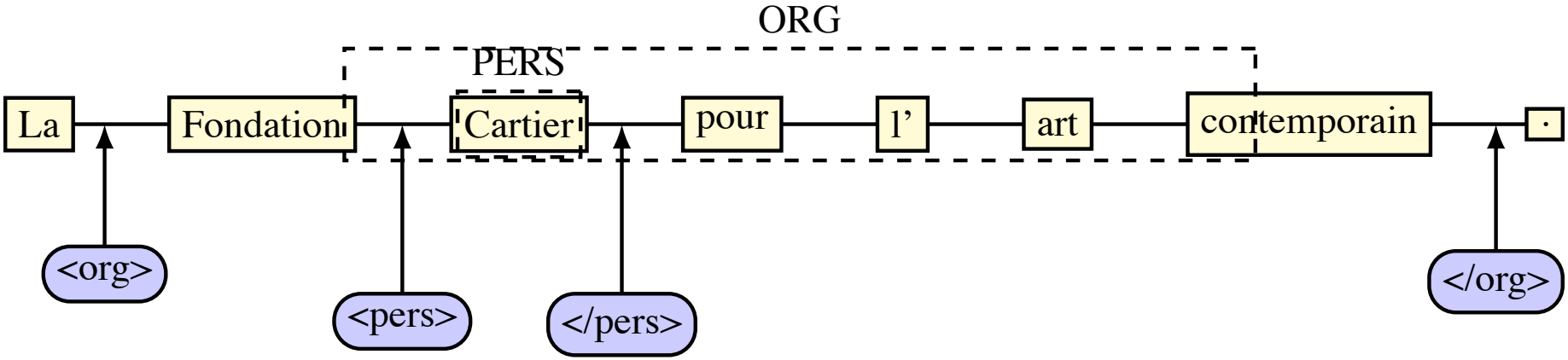
Représentations des annotations



Représentations des annotations

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

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

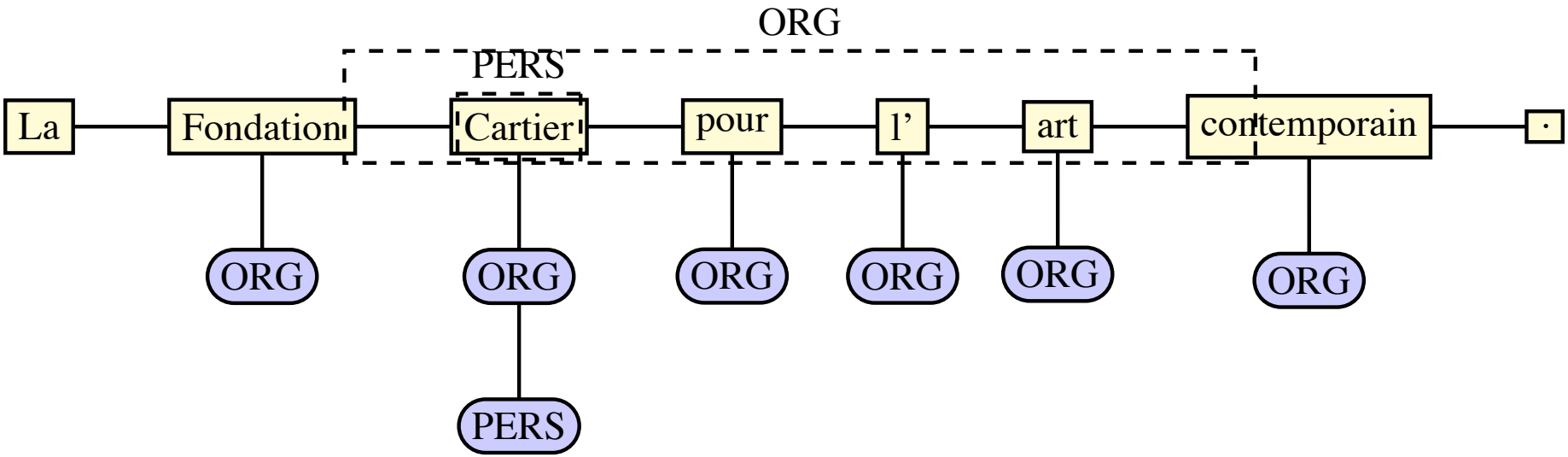


Représentations des annotations

- **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

  ⇒ Modèle pour ajouter un marqueur **individuel** ?

Les marqueurs d'annotation

Annotation en composants

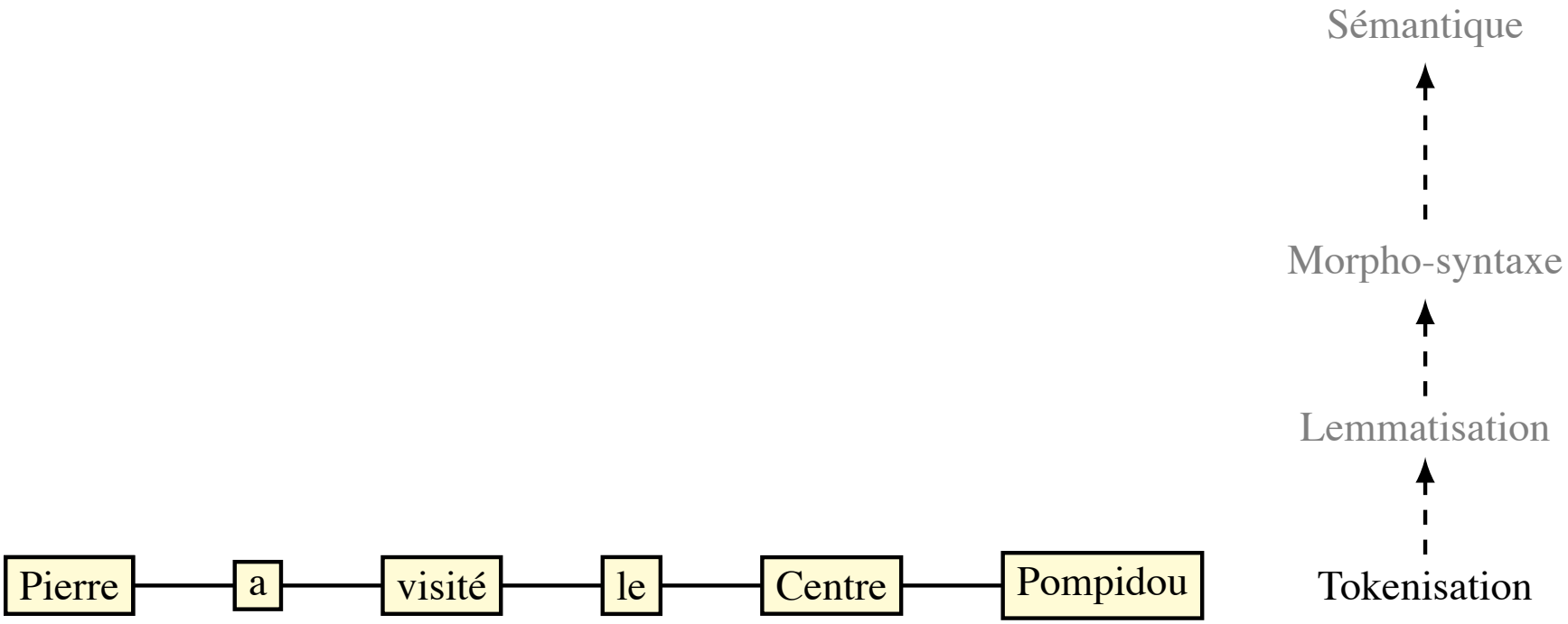
```
` Le <date> <jour> 20 </jour> <mois> mai </mois> <annee>
2010 </annee> </date>, pour <montant> <num> 13 </num>
<monnaie> euros </monnaie> </montant>, il avait visité le
<org> centre <pers> <prenom> Georges </prenom> <nom>
Pompidou </nom> </pers> </org> à <loc> Paris </loc>. '
```

- **Marqueurs** comme instructions de structuration locale
  - Prise en compte du **contexte local** (séquences)
  - Besoin **limité** d'inférences complexes
  - ⇒ Modèle pour ajouter un marqueur **individuel** ?
- Rétrospective des théories linguistiques

⇒ Proposition de propriétés de **stabilité** et d'**opérabilité**

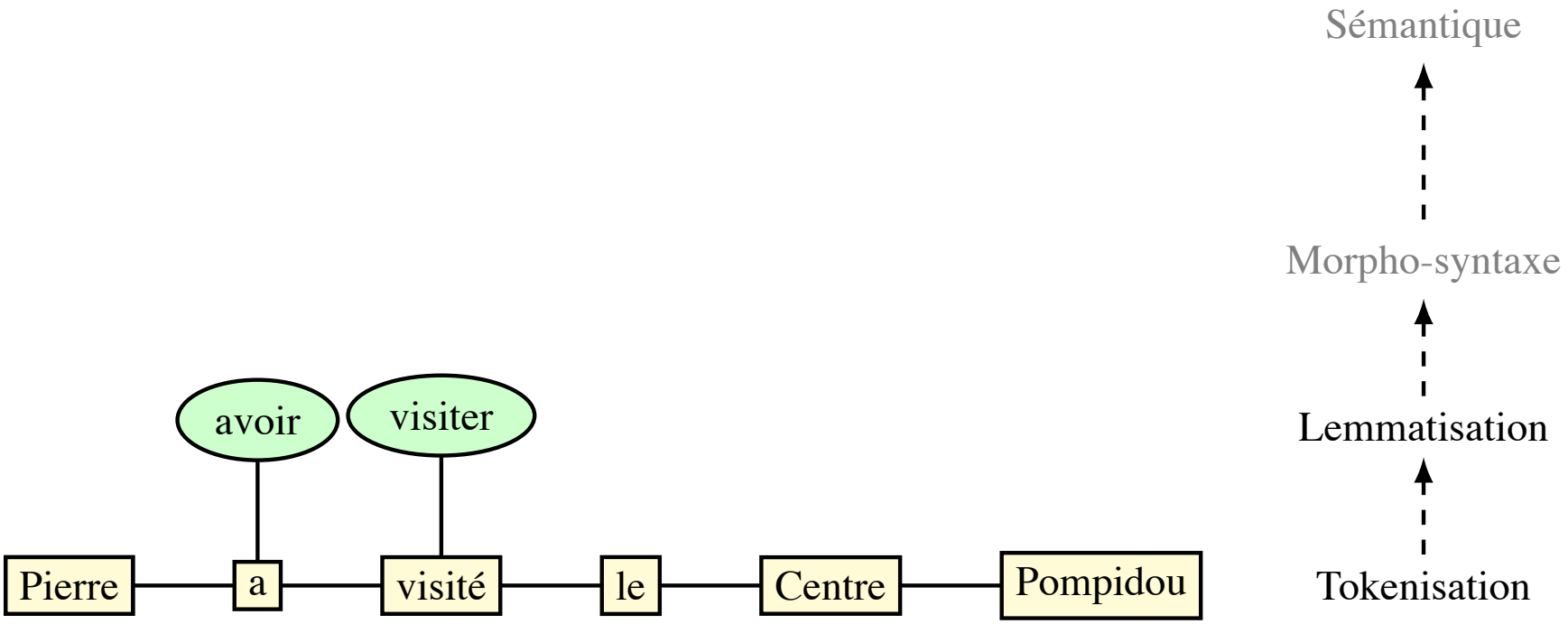
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse



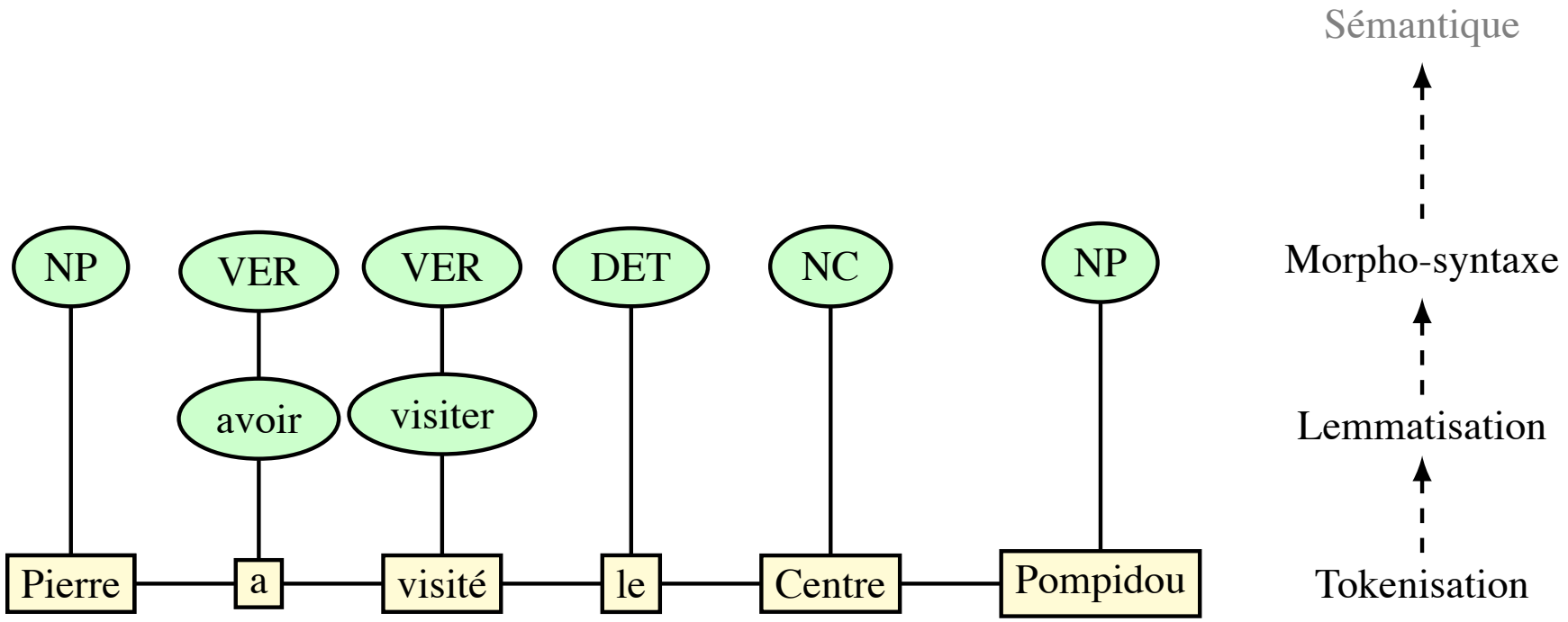
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse



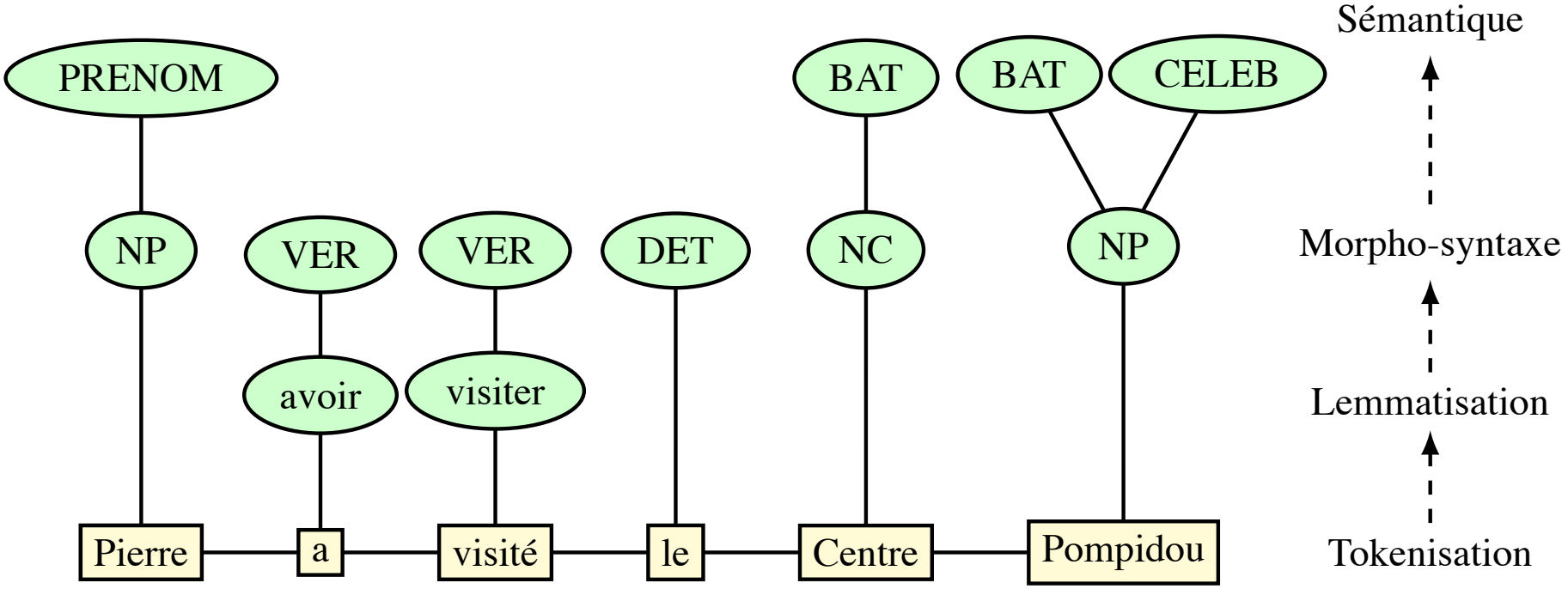
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse



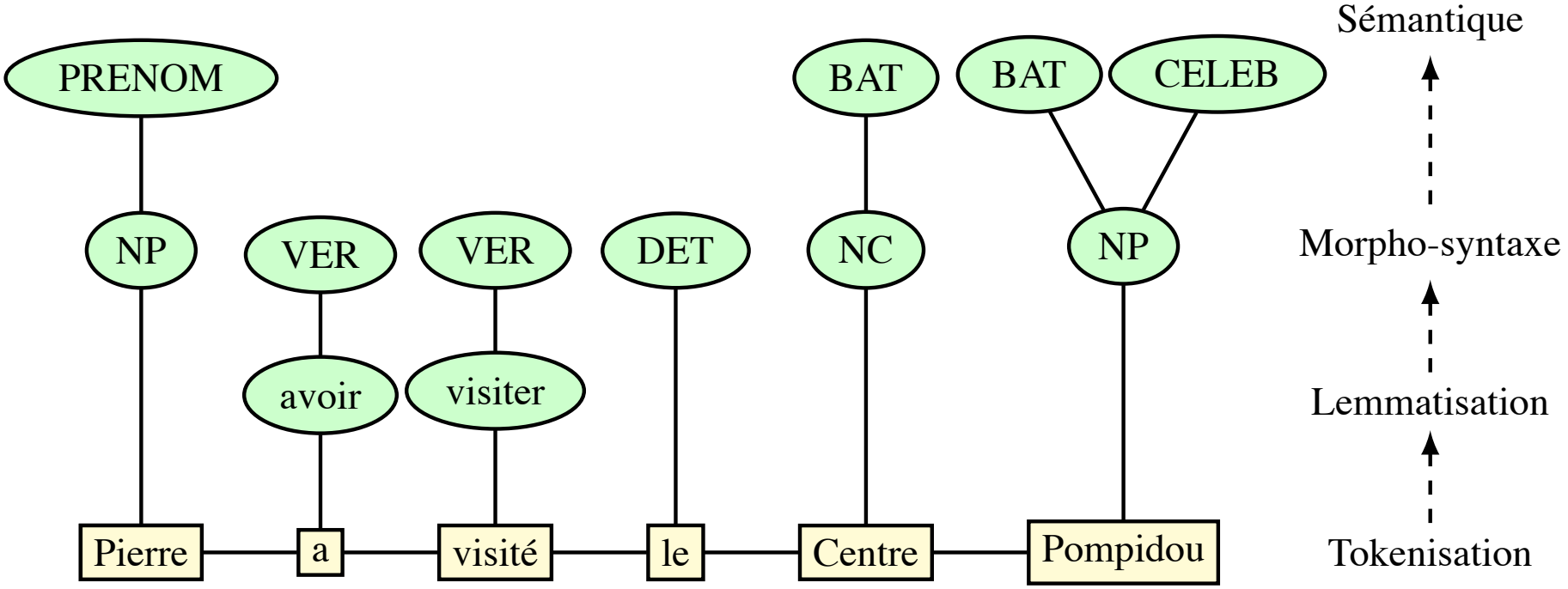
Traiter automatiquement les langues naturelles

- Stratégie incrémentale d'analyse



Traiter automatiquement les langues naturelles

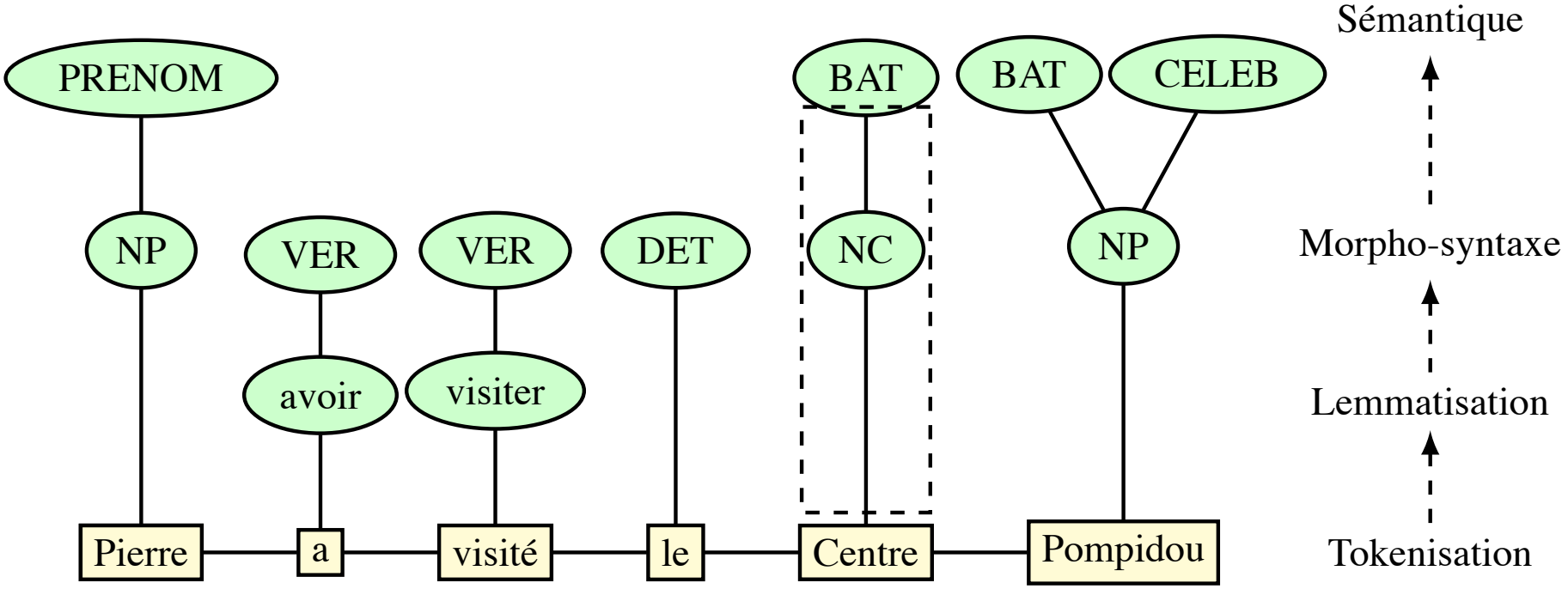
- Stratégie incrémentale d'analyse



- Axes pour les analyses ultérieures

Traiter automatiquement les langues naturelles

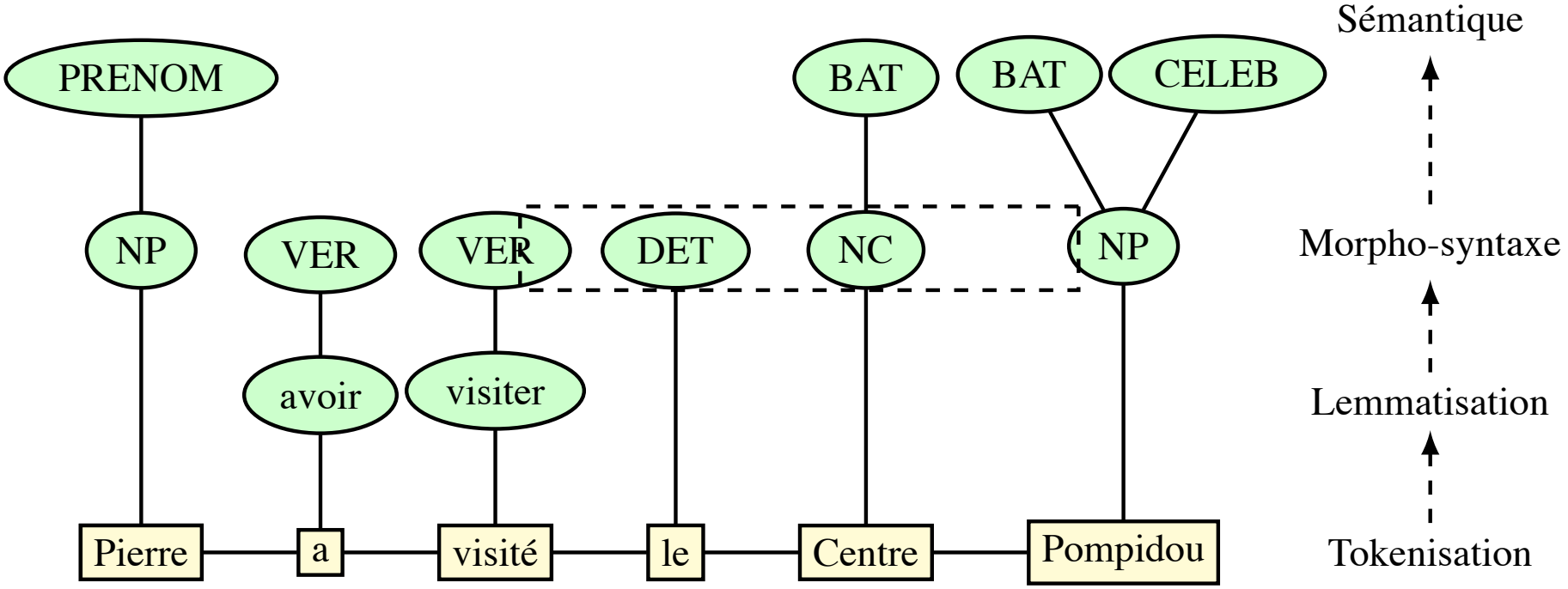
- Stratégie incrémentale d'analyse



- 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

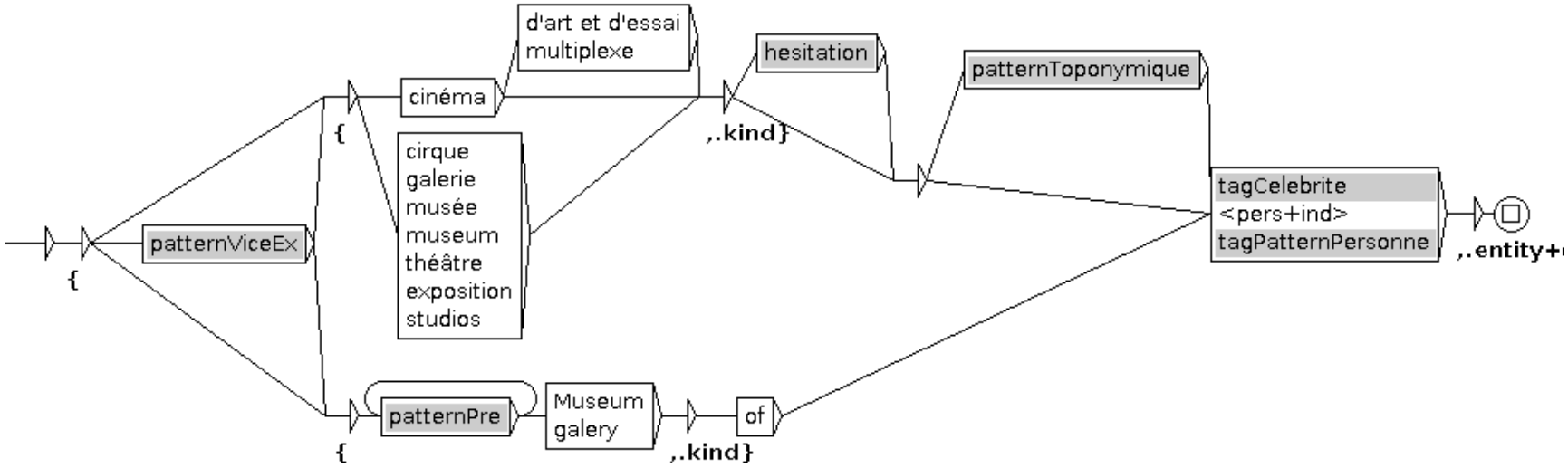
Approches pour la REN

- 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

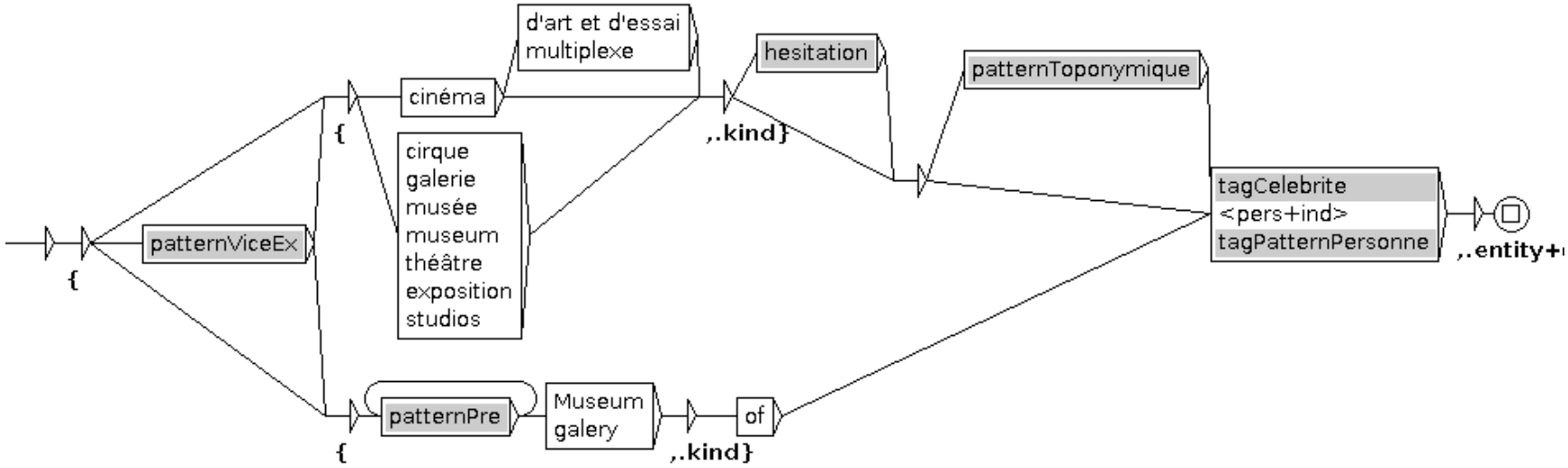
  [McDonald 1996, Friburger 2002, Stern & Sagot 2010, Maurel et.al. 2011]



Approches orientées connaissances

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

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



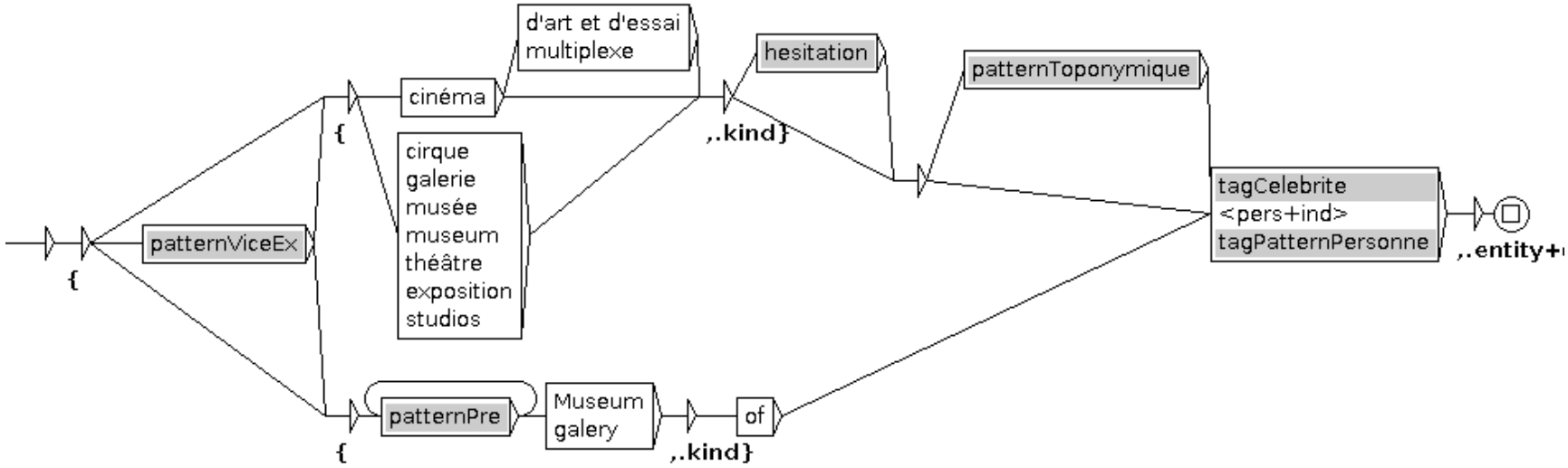
⇒ + Très **précise** (contexte) et **contrôlée**

⇒ + 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]



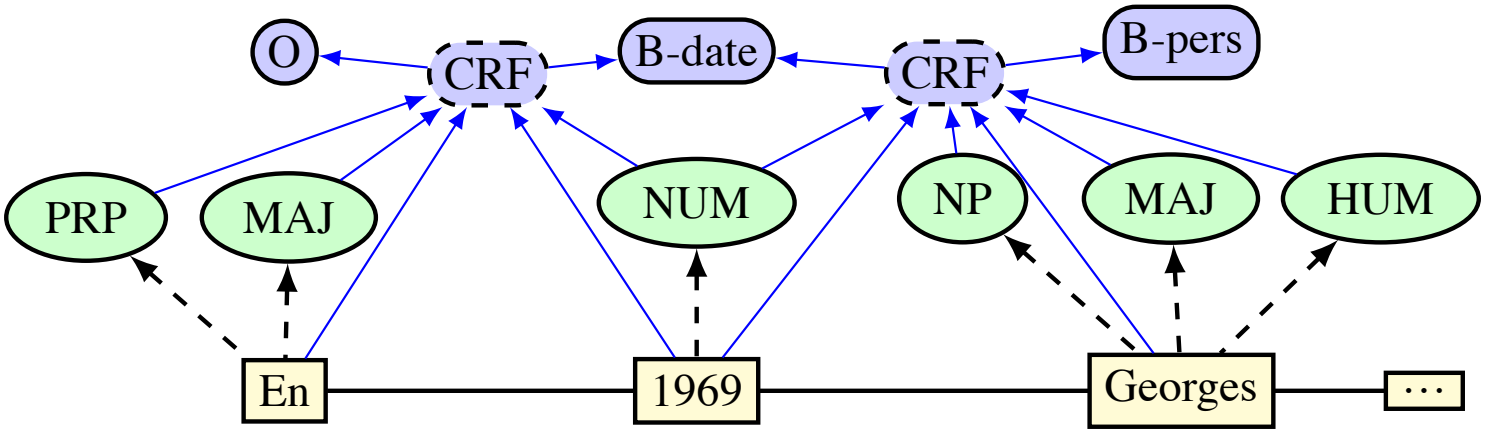
⇒ + Très **précise** (contexte) et **contrôlée**

⇒ + Connaissance capitalisée

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

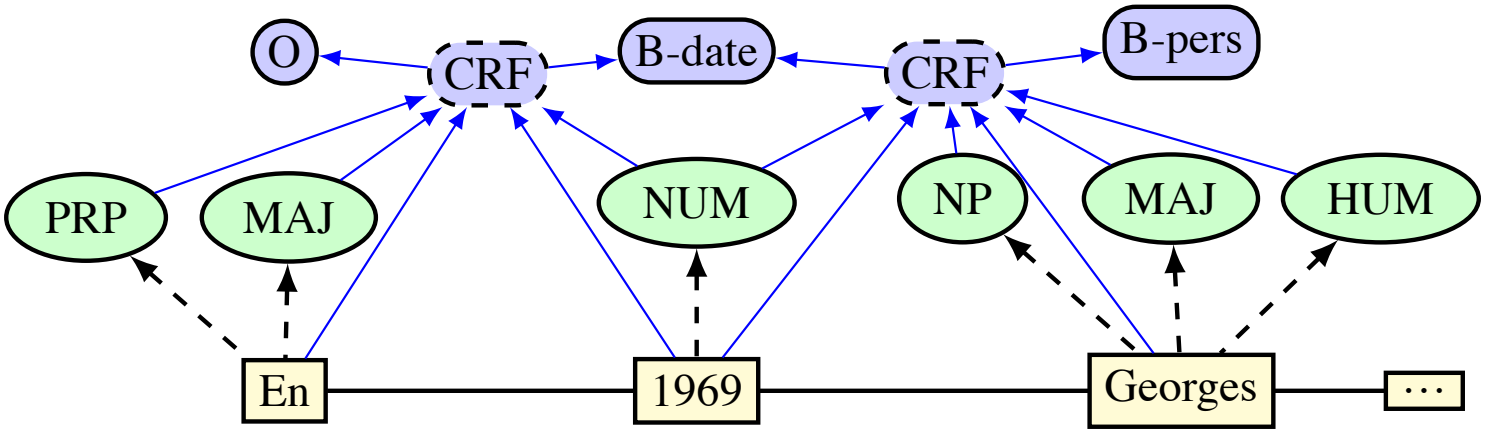
Approches orientées données

- 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

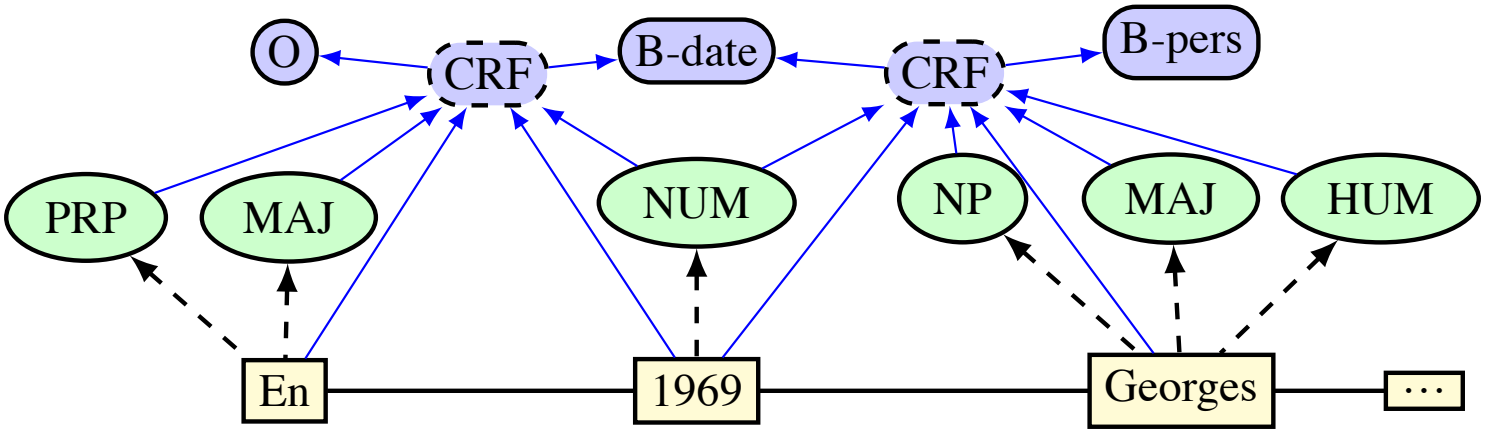
- Estimation de la probabilité pour un mot de recevoir un type d'entité nommée (classification) [Borthwick et. al 1998, Mikheev 1999, McCallum 2000, Raymond & Fayolle 2010]
  - Fonctions caractéristiques (régression logistique)
  - Transitions sur la séquence (HMM)



⇒ + **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

- Par rapport aux approches orientées connaissances
  - ⇒ Extraire des connaissances **complexes** à partir des données
  - ⇒ Analyses préalables **ambigües**

Positionnement

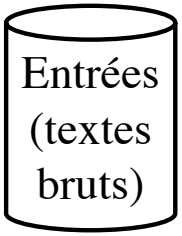
- Par rapport aux approches orientées connaissances
  - ⇒ Extraire des connaissances **complexes** à partir des données
  - ⇒ Analyses préalables **ambigües**
- Par rapport aux approches orientées données
  - ⇒ Ne pas modéliser un processus *"mot-à-mot"* ($\neq$ classification)
  - ⇒ Capacité à générer une **structure** d'annotation

Positionnement

- Par rapport aux approches orientées connaissances
  - ⇒ Extraire des connaissances **complexes** à partir des données
  - ⇒ Analyses préalables **ambigües**
- Par rapport aux approches orientées données
  - ⇒ Ne pas modéliser un processus *"mot-à-mot"* ($\neq$ classification)
  - ⇒ Capacité à générer une **structure** d'annotation

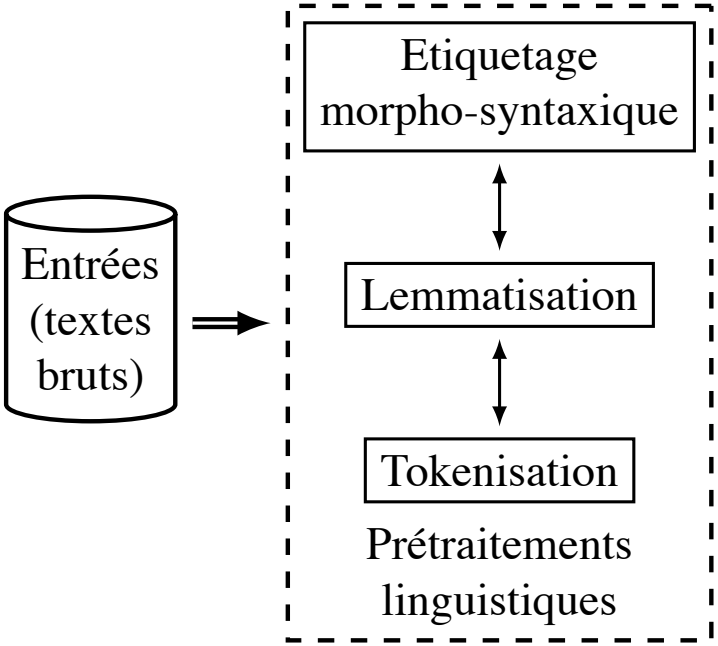
⇒ Originalité : rechercher séparément les **marqueurs** qui délimitent le **début** ou la **fin** d'une entité nommée

Cadre expérimental



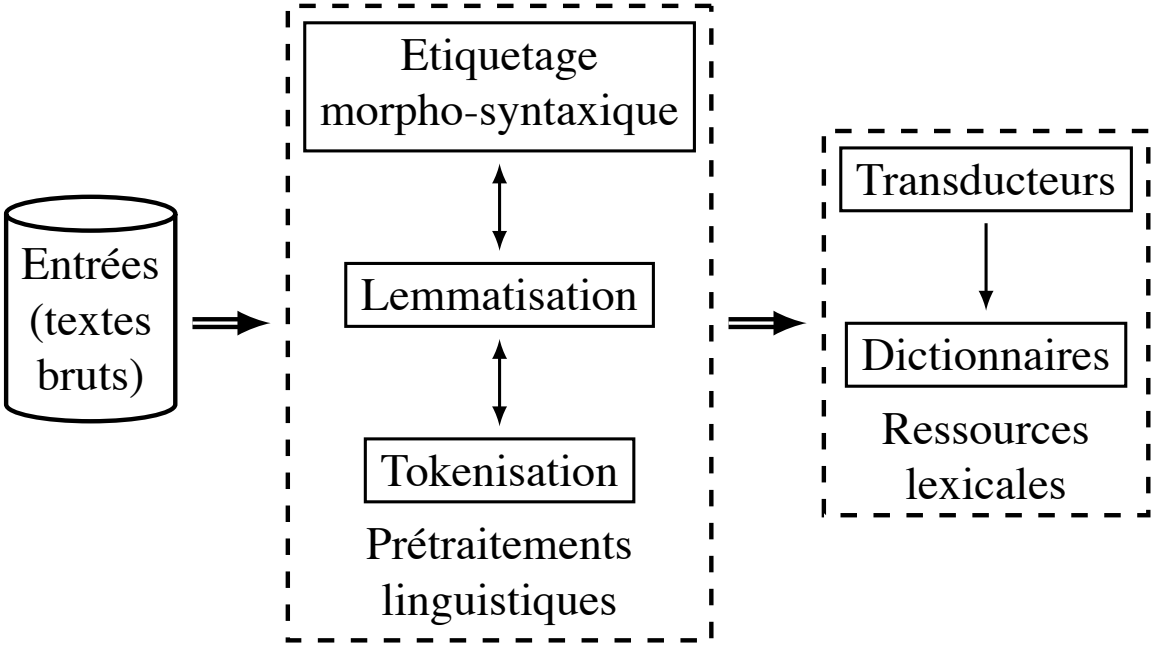
Cadre expérimental

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



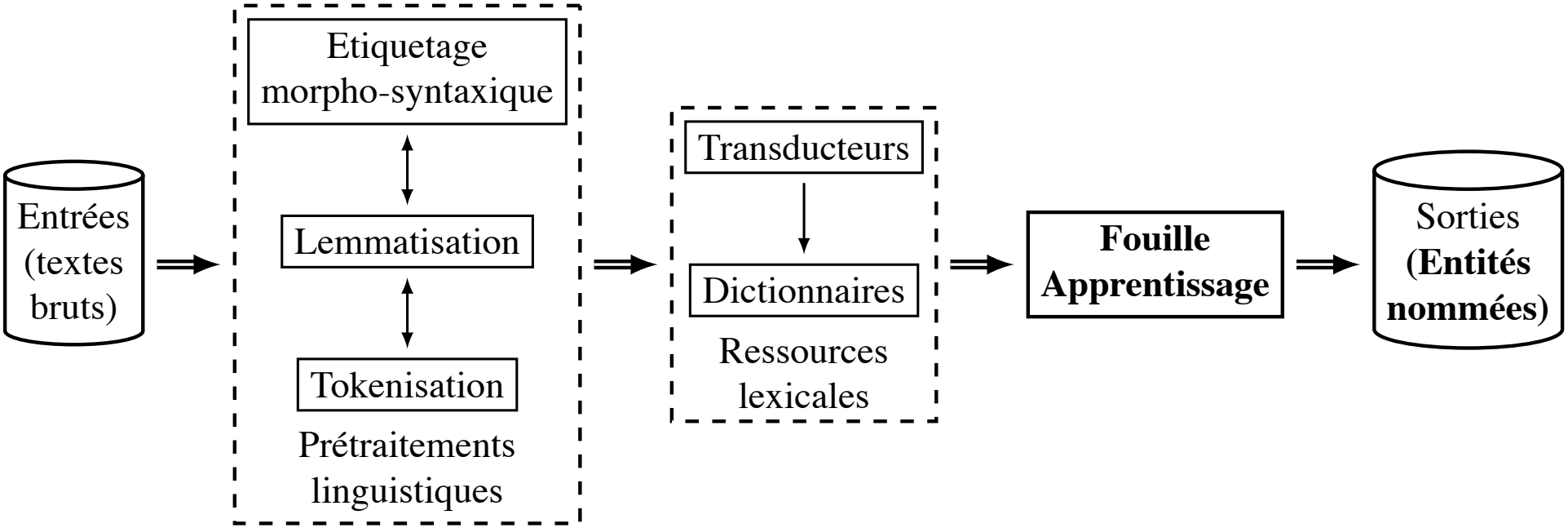
Cadre expérimental

- 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)



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]

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]
- 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]
- 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$\oplus$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

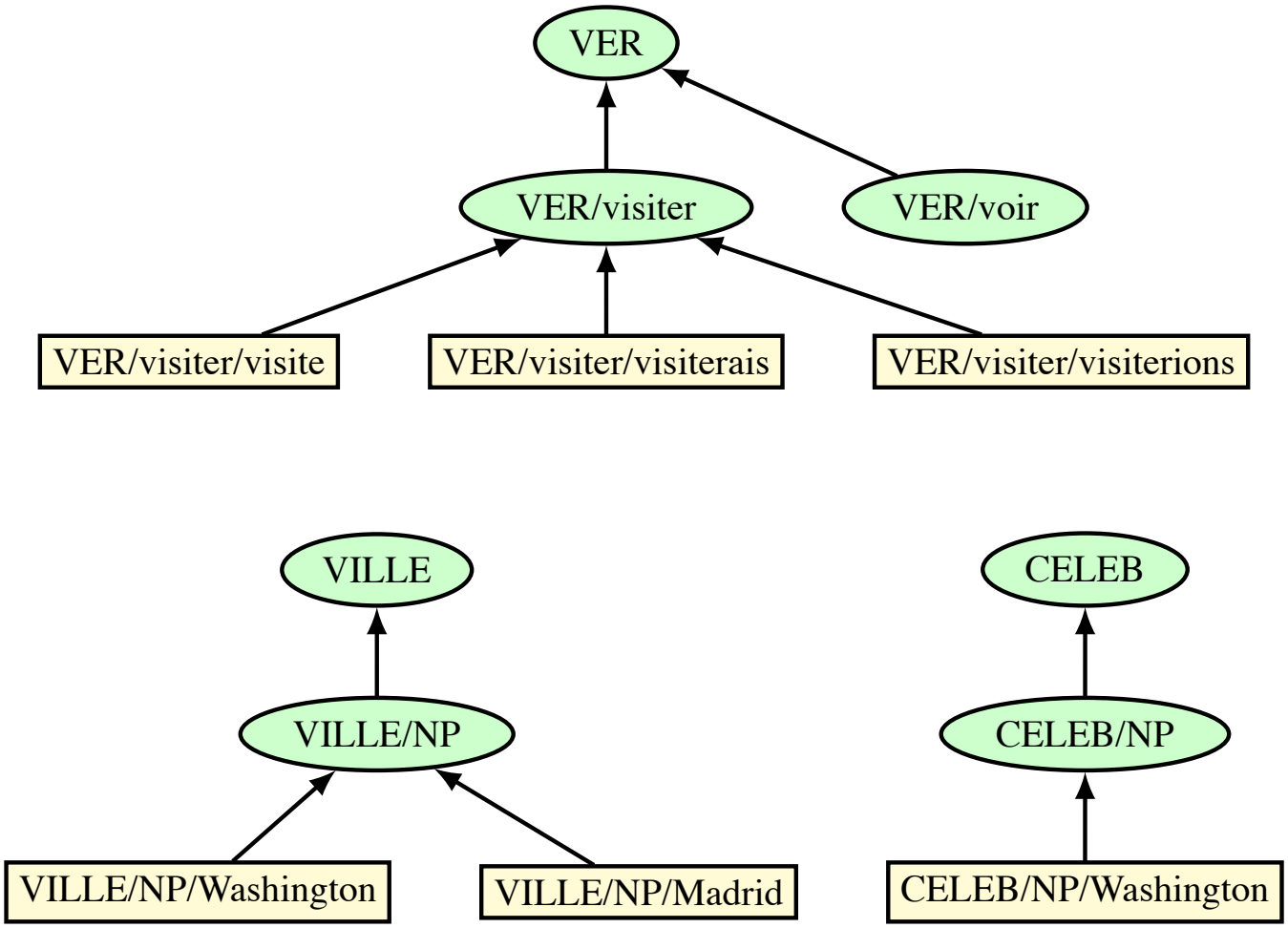
$E =$ `Il visite Washington .'

$R(E) =$ `PRO/il/Il VER/visiter/visite VILLE/NP/Washington$\oplus$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



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

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'
      ⇒ `DET/Le/le <date> NUM juillet NUM </date>'
```

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'
      ⇒ `DET/Le/le <date> NUM juillet NUM </date>'
```

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

Critères pour extraire les règles

Critères pour extraire les règles

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

Critères pour extraire les règles

⇒ **Filtrer** les règles qui ont a priori peu d'intérêt

⇒ 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

⇒ **Filtrer** les règles qui ont a priori peu d'intérêt

⇒ 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

⇒ **Filtrer** les règles qui ont a priori peu d'intérêt

⇒ 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

  ⇒ Couvrent les mêmes exemples

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>'

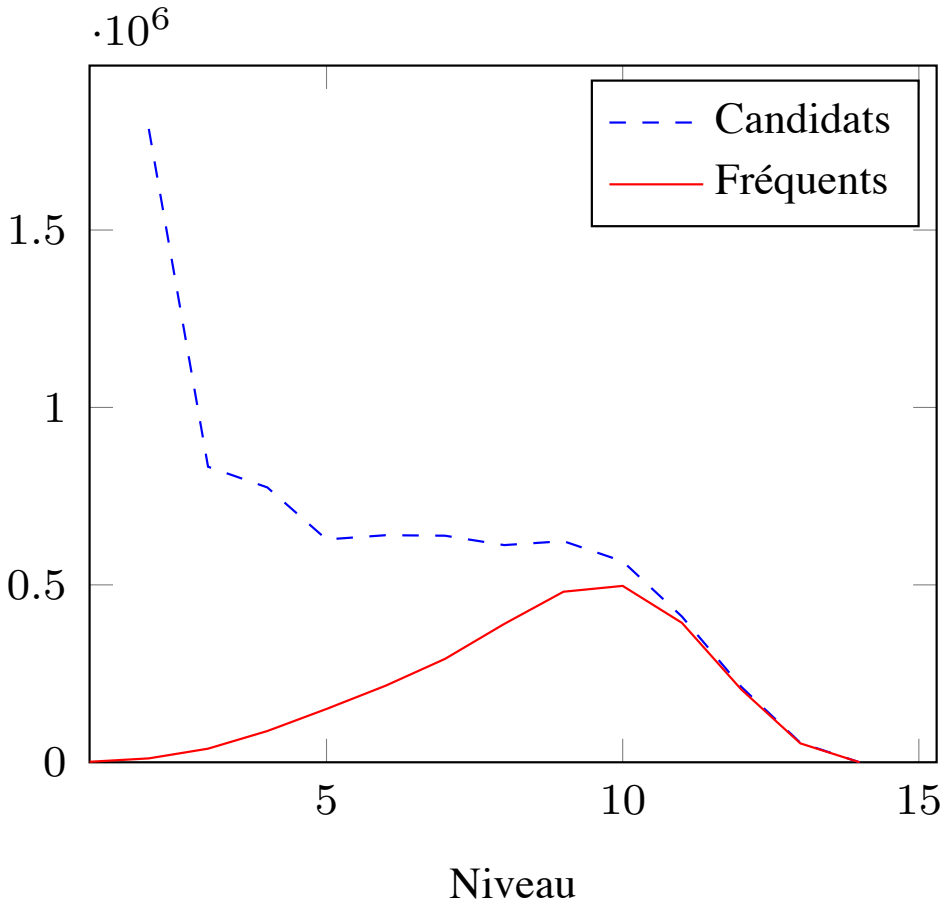
...(total : 26 motifs)

Extraction des motifs

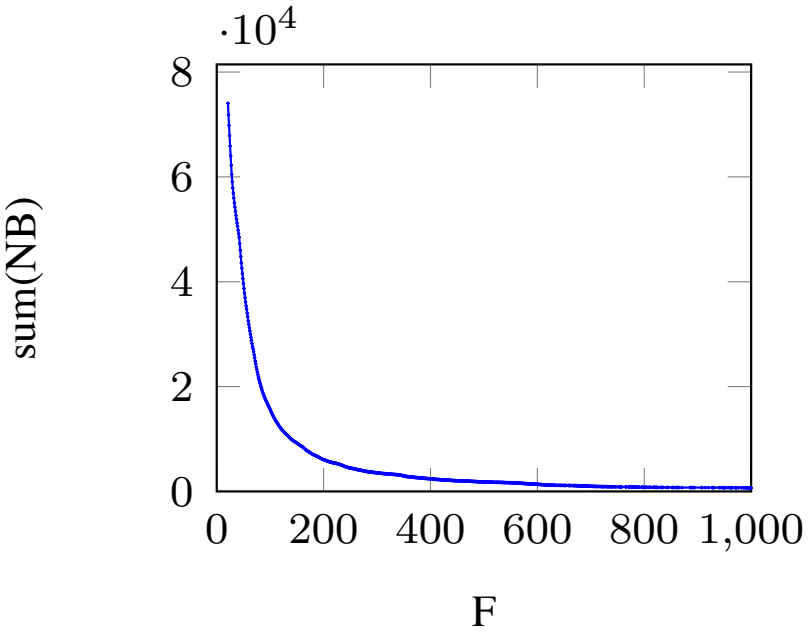
- Algorithme par niveaux (avec optimisations)

Extraction des motifs

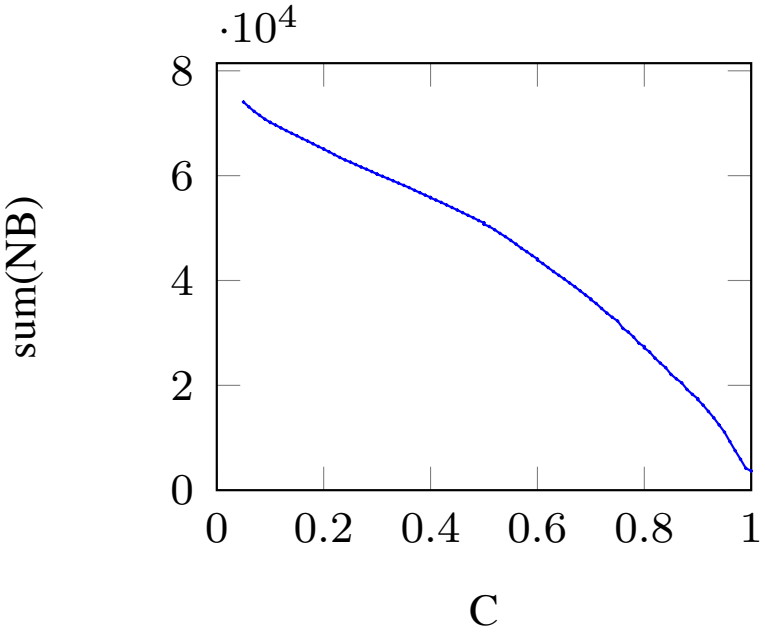
- Algorithme par niveaux (avec optimisations)



Nombre de motifs extraits

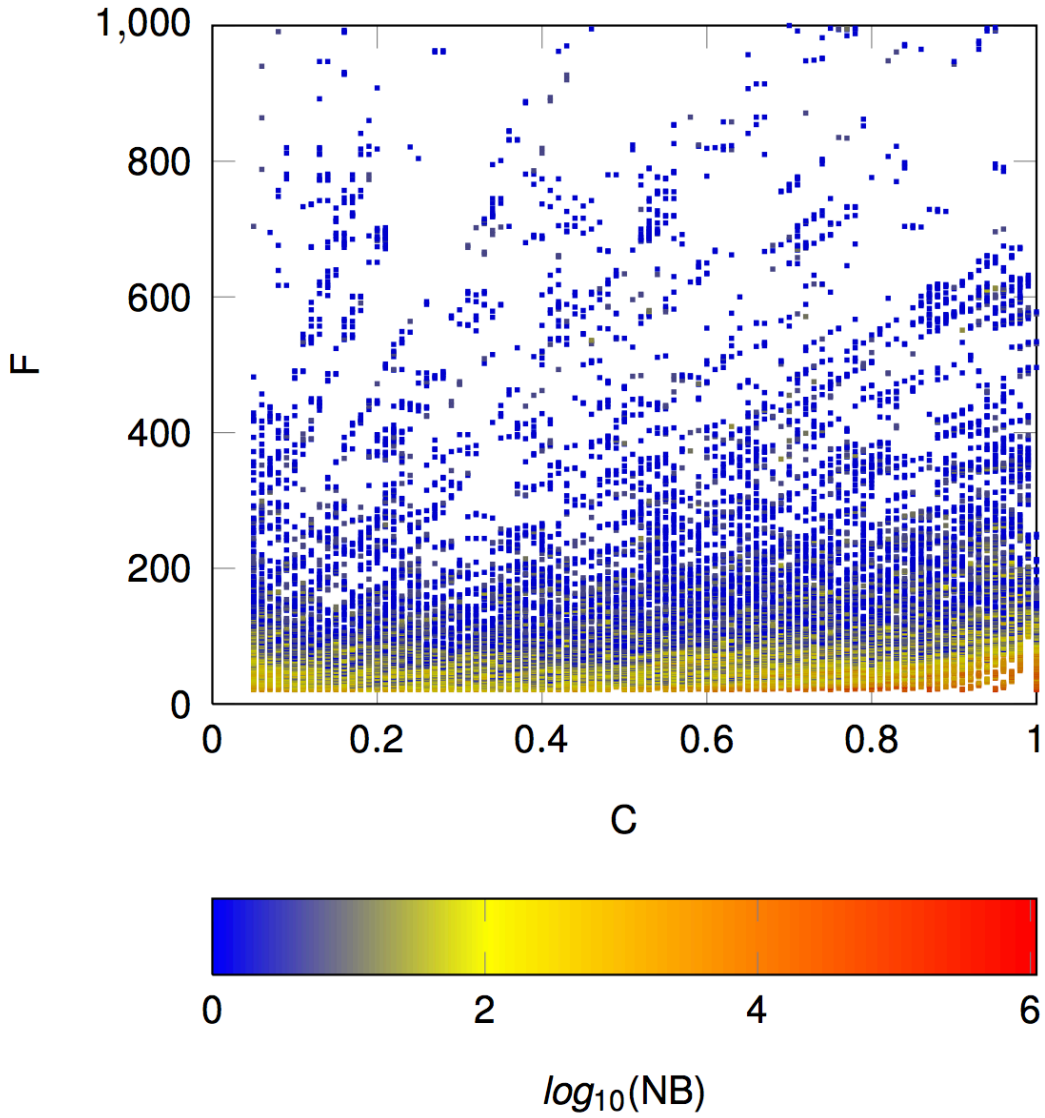


(d) Fréquence



(e) Confiance

Spectre des motifs extraits



Modèles de marqueurs

- Pondération de règles
  - Modèle probabiliste sur marqueurs (régression logistique, Scikit)
  - **Paramétrage itératif orienté données**

Modèles de marqueurs

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

$\Rightarrow$ Calcul des marqueurs les plus probables

Modèles de marqueurs

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

$\Rightarrow$ Calcul des marqueurs les plus probables

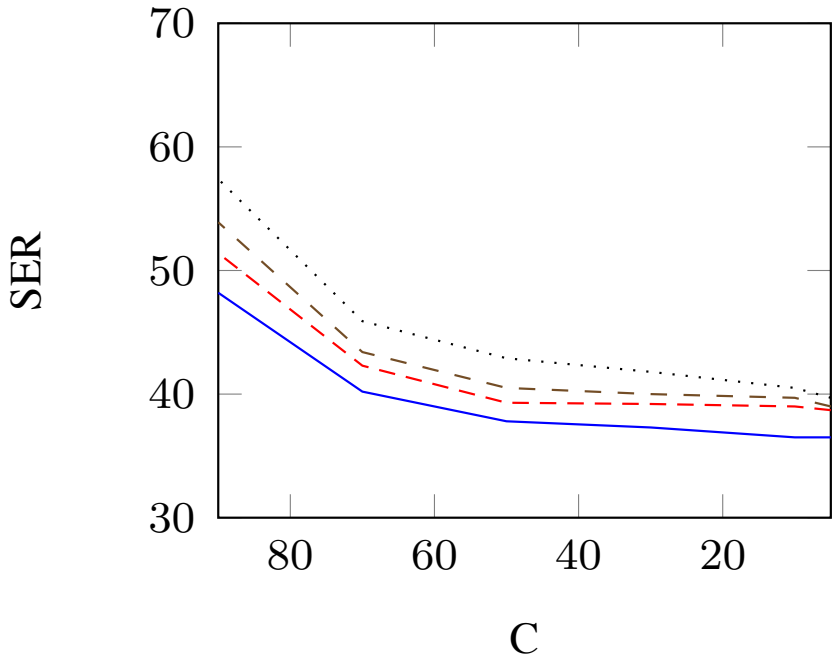
- Annotation structurée : séquences de marqueurs

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

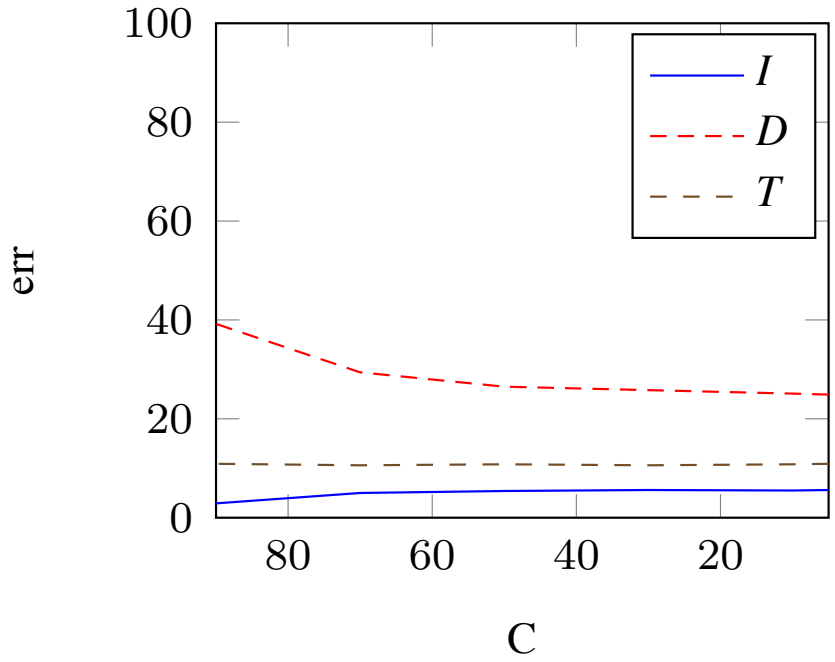
$\Rightarrow$ Calcul de l'annotation la plus probable :

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

Performances globales



(f) Performances

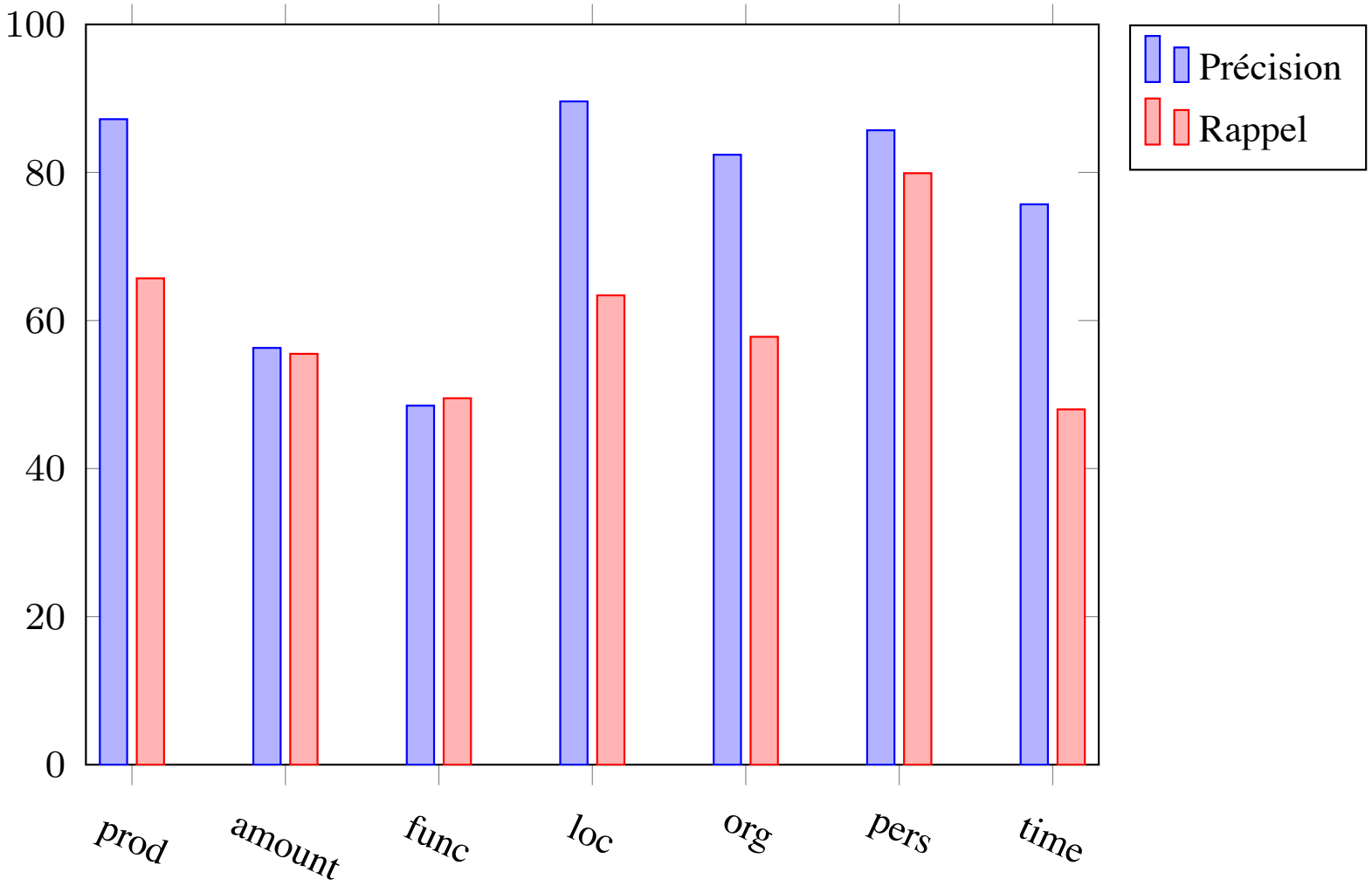


(g) Erreurs, $F \geq 5.10^{-5}$

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, systèmes orientés connaissances / 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

Résoudre /lier les entités (nommées)

- Nouvelles perspectives pour les entités (nommées)
  - Reconnaissance insuffisante : *F. Mitterand*
  - Liaison avec le **web sémantique** (LOD)

  $\Rightarrow$ Reconnaissance de **mentions** d'entités

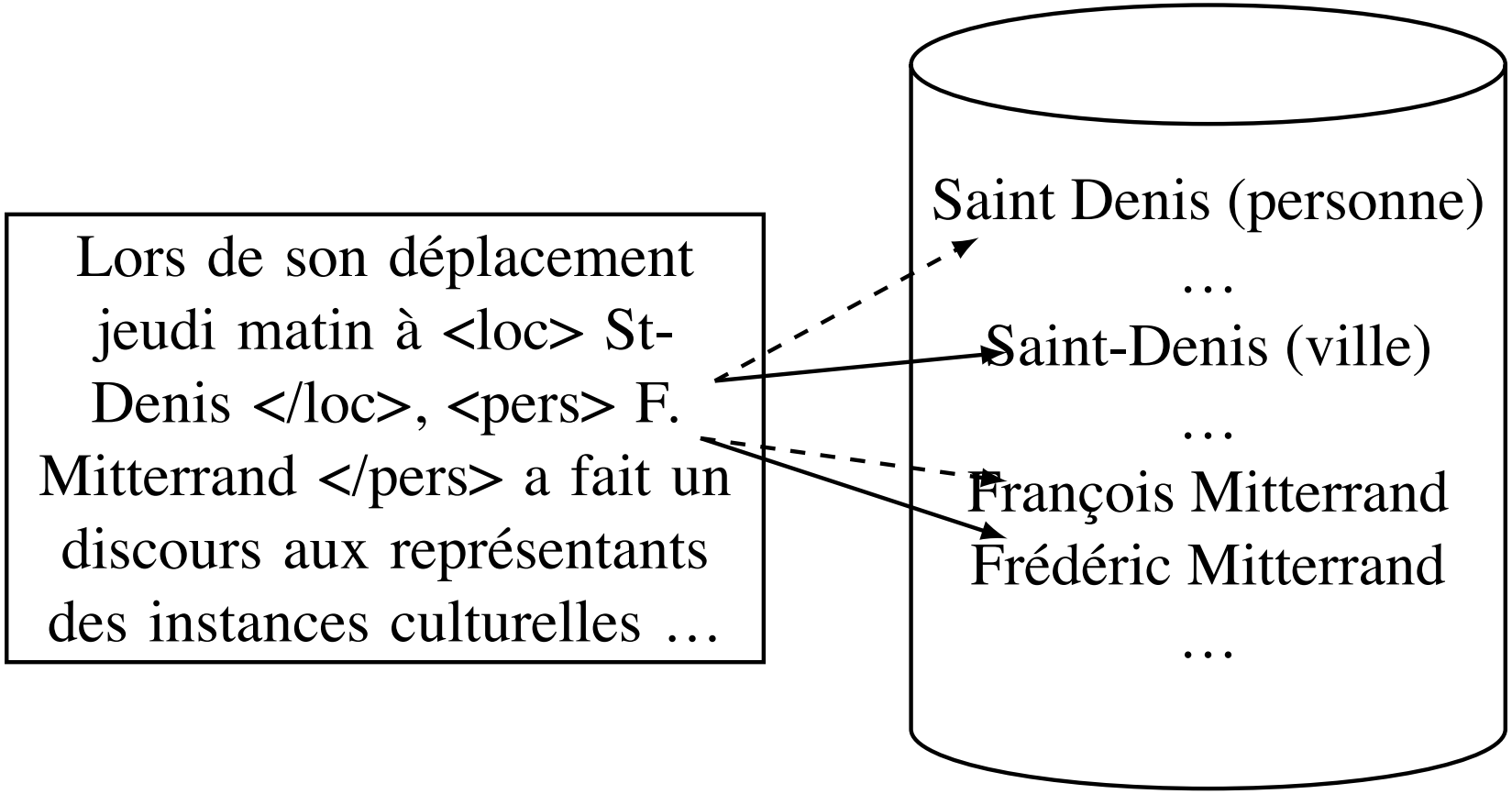
Résoudre /lier les entités (nommées)

- Nouvelles perspectives pour les entités (nommées)
  - Reconnaissance insuffisante : *F. Mitterand*
  - Liaison avec le **web sémantique** (LOD)

  ⇒ Reconnaissance de **mentions** d'entités
- Exploitation d'un **modèle** / référentiel
  - Liste aussi large que possible d'entités d'intérêt
  - Nécessairement **incomplet** (`NIL`)
  - Liaison : coréférences, anaphores, clustering …

⇒ Exploitation intensive de **Wikipedia**

Lier le texte aux bases de connaissances



Principes généraux de résolution

- Prétraitement par **reconnaissance** rudimentaire
  - Insensibilité à la casse (majuscules / minuscules)
  - Variantes liées aux jeux de caractères (diacritiques, liaisons, etc.)
  - Suppression des mots-outils / de parenthèses
  - Génération automatique d'acronymes à partir des formes

$\Rightarrow$ De nombreuses liaisons potentielles

Principes généraux de résolution

- Prétraitement par **reconnaissance** rudimentaire
  - Insensibilité à la casse (majuscules / minuscules)
  - Variantes liées aux jeux de caractères (diacritiques, liaisons, etc.)
  - Suppression des mots-outils / de parenthèses
  - Génération automatique d'acronymes à partir des formes

$\Rightarrow$ De nombreuses liaisons potentielles

- La **désambiguisation** contextuelle
  - **Texte** du document : mots, domaine, autre entités (anaphores)
  - **Popularité** du référent
  - Attributs du référent : type, nom / titre (redirections), description

$\Rightarrow$ Apprentissage (SVM) et classification des référents par mention

Quelques résultats de résolution

- Performances (anglais, textes généraux) : 80% à 90%

Quelques résultats de résolution

- Performances (anglais, textes généraux) : 80% à 90%
- Nombreux logiciels en ligne, dont
  - BabelFy (BabelNet)
  - Wikipedia Spotlihght
  - Alchemy
  - The Wiki Machine
  - NERD
  - …

$\Rightarrow$ Peu de résolveurs par domaine (historique, biomédical, juridique, finance, etc.)

Plan

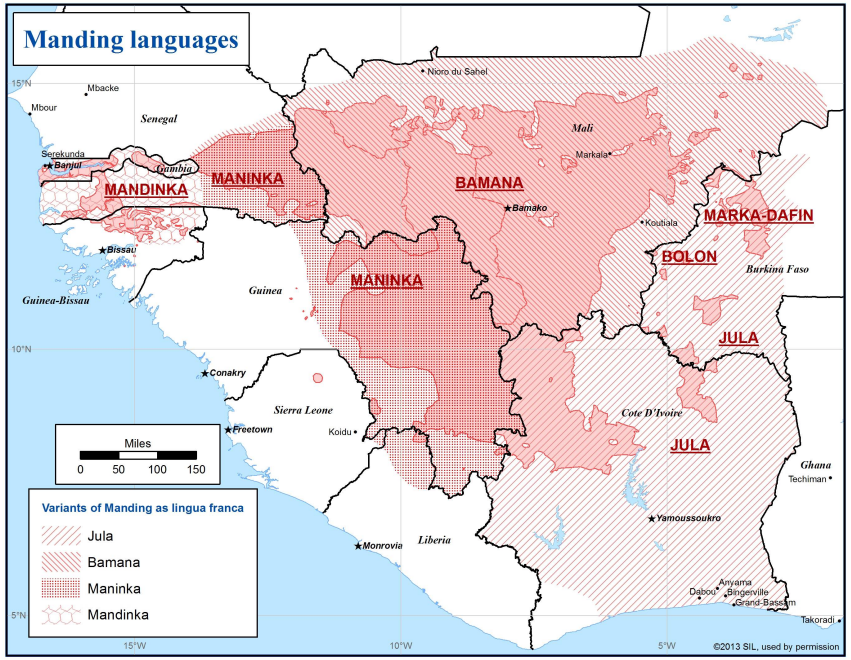
1. La désambiguisation lexicale : des formes aux sens

2. Reconnaissance et liaison des entités nommées

3. Enrichissement lexical d'un corpus bambara
   - Contexte
   - Projet MANTAL
   - Annotation morpho-syntaxique
   - Tonalisation du bambara
   - Conclusions et perspectives

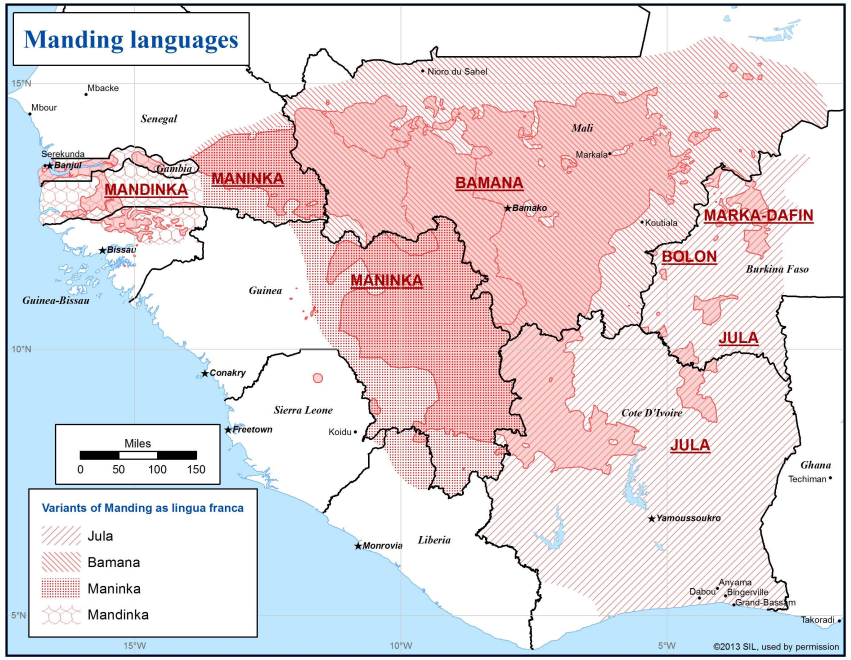
4. Quelques travaux préliminaires sur l'arabe

Contexte



- Parlée principalement au Mali (diglossie : français, 20%)
- Ou « bamanaka »

Contexte



- Parlée principalement au Mali (diglossie : français, 20%)
- Ou « bamanaka »
- 4M de locuteurs (10M en 2 ème langue)
- Véhiculaire, tradition orale
- Macro-langue mandingue (avec dioula, malinké, etc.)

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

⇒ Langues peu standardisées

Quelques considérations linguistiques

- Prononciation : 7 voyelles, 20 consonnes, 3 tons
- Alphabet
  - Langues mandingues : alphabet n'ko (1950, Unicode 5.0, rtl)
  - Bambara : alphabet latin
  - Depuis 1982 : ɛ (U+025B/U+03B5), ɔ, ŋ, ɲ

⇒ Langues peu standardisées

- Grammaire
  - Type : S AUX O V X, tonale
  - Pas de genre grammatical
  - Pas de conjugaison (marques prédicatives AUX)
  - Peu de flexion (-w : pluriel)

Travaux antérieurs

- Références sur le bambara
  - (Binger, 1886) Essai sur la langue bambara.
  - (Sauvant, 1926) Dictionnaire bambara-français.
  - (Vydrin, 1999) Les parties du discours en bambara.
  - (Vydrin, 1999) Manding-english dictionary (maninka, bamana).
  - (Dumestre, 2003) Grammaire fondamentale du bambara.
  - (Bailleul, 2007) Dictionnaire bambara-français.
  - (Vydrin, 2008) Glossed electronic corpora of Mande languages.
  - (Dumestre, 2011) Dictionnaire bambara-français.
  - (Vydrin, 2011) Corpus électronique annoté des textes bambara.
  - (Enguehard, 2012) Vers l'info. de langues d'Afrique de l'Ouest.
  - (Maslinsky, 2014) Daba : a model and tools for manding corpora.

Corpus bambara de référence

- Collecte de textes en bambara
  - Publiés (périodiques, littérature) ou non (lettres, trans.)
  - Normalisation des textes (orthographe, tons, etc.) ?
  - Textes en ligne `http://cormand.huma-num.fr/biblio/`

⇒ Volume : 2,3M mots

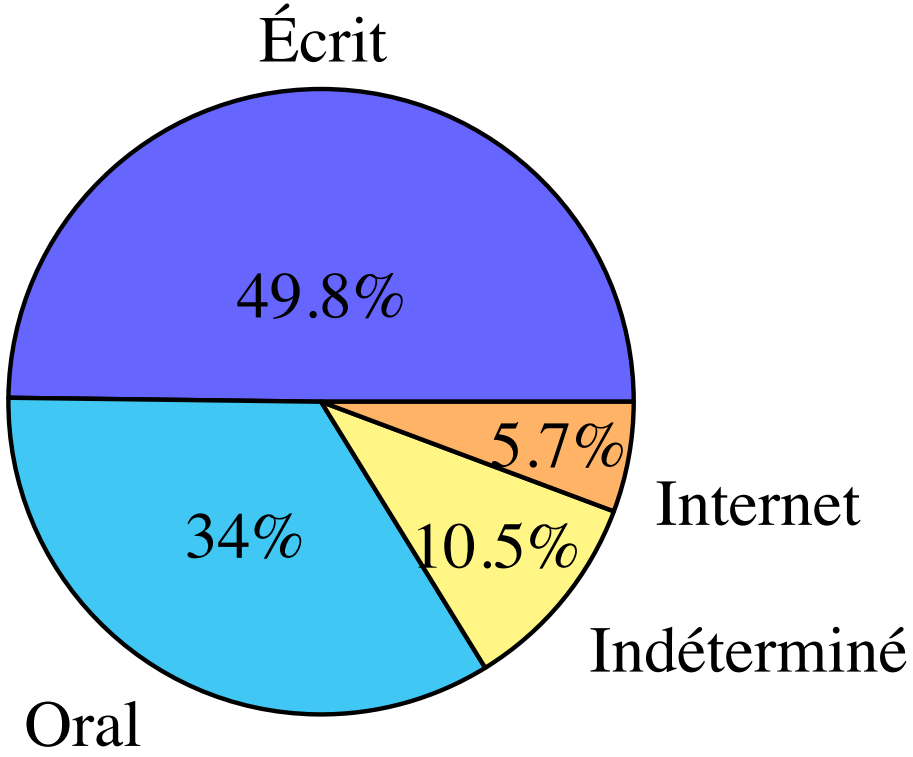
- Utilisation essentiellement linguistique
  - Apprentissage de la langue
  - Études linguistiques sur corpus
  - Annotation (parties du discours, lemmes, morphologie, gloses)

⇒ Labex EFL, axe 6 : ressources linguistiques

⇒ Site internet : `http://cormand.huma-num.fr/`

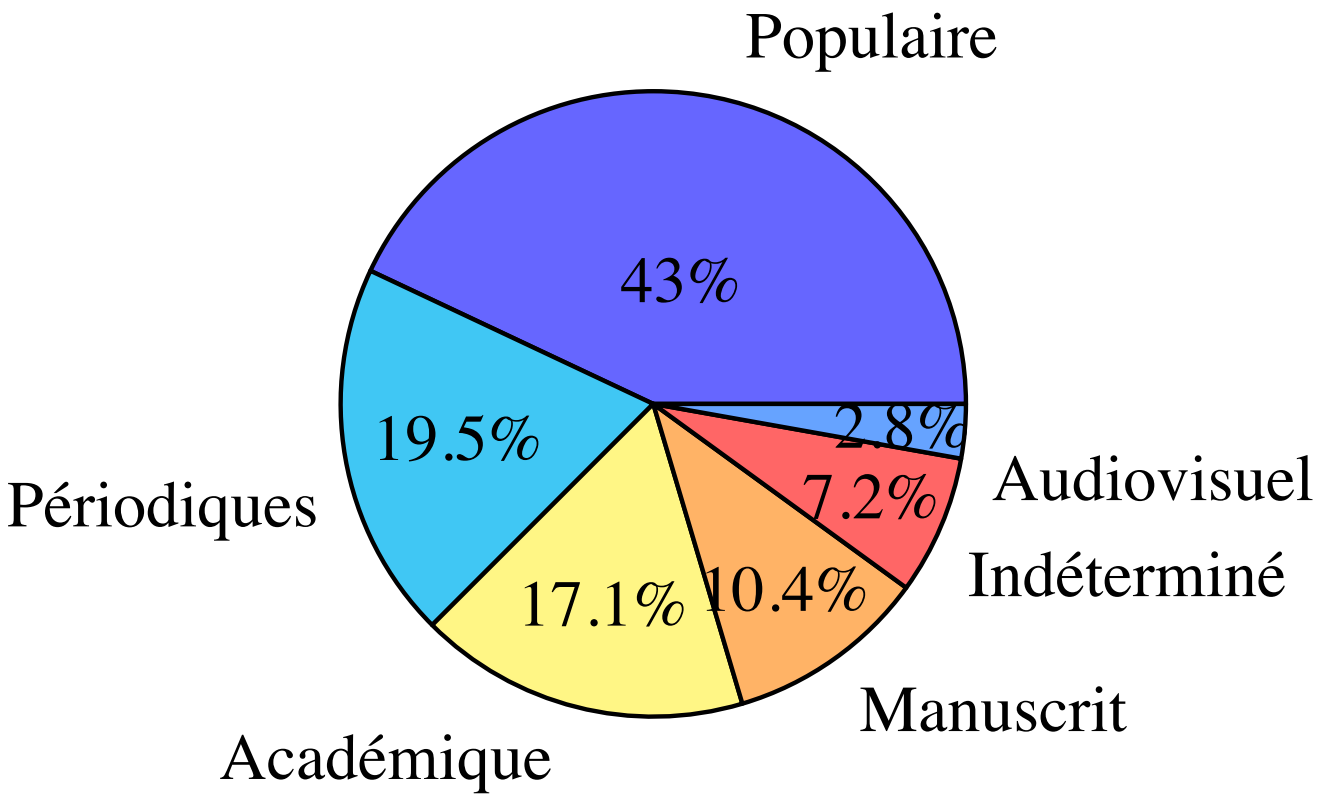
⇒ Modélisation linguistique informatisée (SketchEngine)

Médiums du corpus



⇒ Prédominance de l'écrit, mais souvent issu de l'oral (contes, etc.)

Sources du corpus



Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

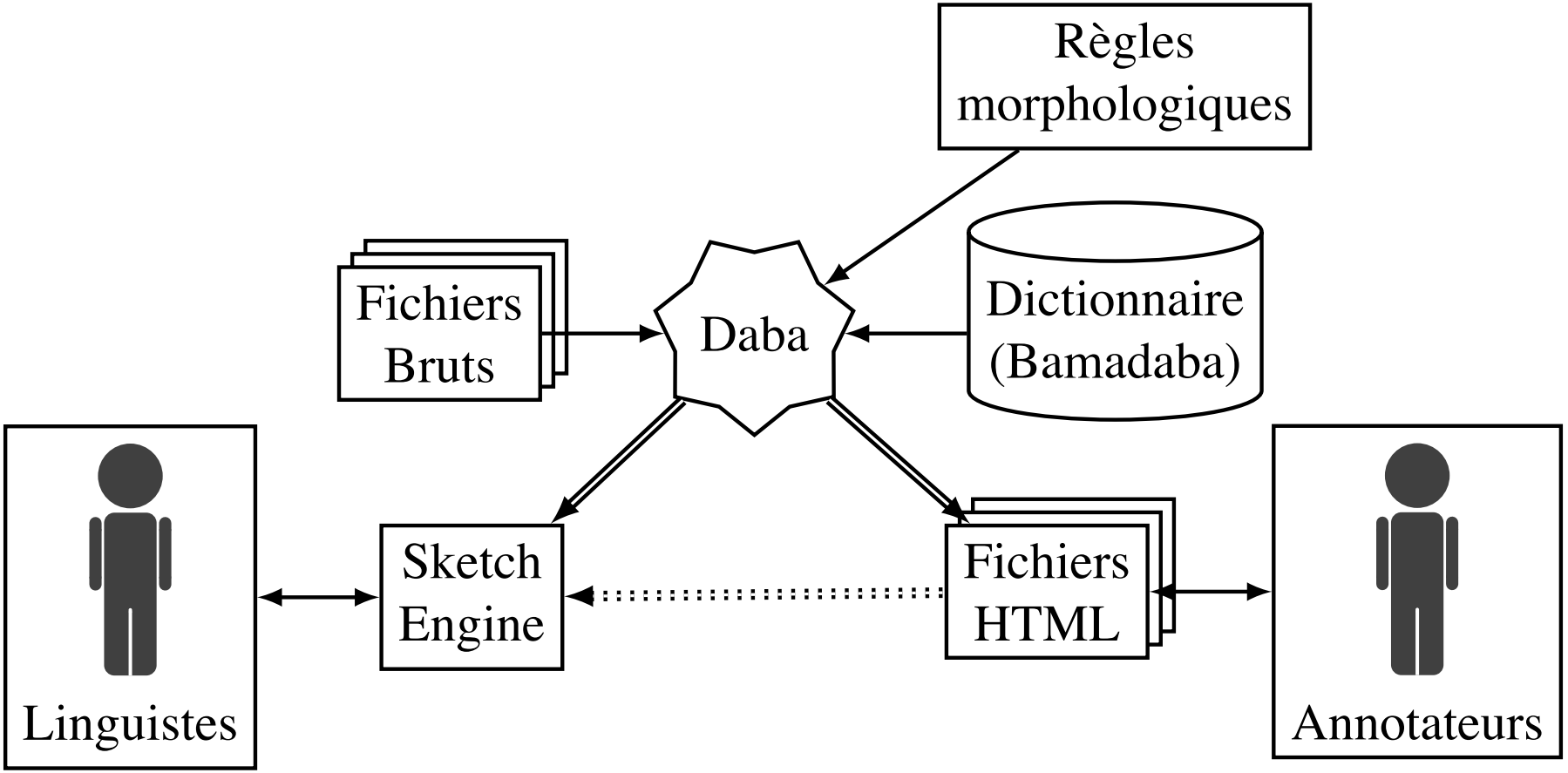
Annotation du corpus

- Utilisation de **Daba** (Maslinsky, 2014)
  - Tokenisation
  - Recherche dans les dictionnaires
  - Analyses morphologiques

$\Rightarrow$ Pré-annotation automatique et ambiguë

- Annotation manuelle
  - Peu de moyens (profs, étudiants, bénévoles)
  - Validation humaine
  - Correction / normalisation

Schéma de fonctionnement



Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - (Entités nommées)
  - (Malinké)

$\Rightarrow$ Exploitation des données annotées (supervision)

Objectifs de MANTAL

- Traitement des langues MANdingues avec des outils TAL
  - Normalisation / uniformisation
  - Parties du discours
  - (Entités nommées)
  - (Malinké)

⇒ Exploitation des données annotées (supervision)

- Cadre général
  - Collaboration LLACAN / ERTIM
  - Durée : 3 ans
  - Budget interne Inalco (stages, prestations, missions)

⇒ Interactions avec les linguistes

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement
  - Fichier source

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique

    $\Rightarrow$ Automatique, sortie ambiguë

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique
    - ⇒ Automatique, sortie ambiguë
  - Annotation par les linguistes
  - ⇒ Ambiguïtés résiduelles

Procédures d'annotation

- Stage M2 « Annotation automatique en partie du discours en Bambara » (D. Auffret, INALCO)
- Plusieurs niveaux de traitement
  - Fichier source
  - Recherche des ponctuations, nombres et noms propres
  - Utilisation de Daba
    - Recherche dans les dictionnaires
    - Analyse morphologique
    - ⇒ Automatique, sortie ambiguë
  - Annotation par les linguistes
  - ⇒ Ambiguïtés résiduelles

⇒ Plusieurs versions de fichiers à synchroniser

Fichier source

<h>Diɲɛ Yaalala</h>
Nsiirin, nsiirin. N y'a bila den dɔ le kan.
Den nin ye sira dali a facɛ fɛ, k'a b'a fɛ ka taga diɲɛ yaala ka dɔ fara a hakili kan. A facɛ ye sira d'a ma a ka taga yaala. A tagara yaala kosɛbɛ.
A fɔlɔla ka taga ben sogosu dɔ ma. A yɔrɔ bɛɛ tolira, a ko : "E ! Ala bɛ se." Sogo nin wulila ka kum'a fɛ k'a kan'a fɔ ko Ala bɛ se, k'Ala ka se b'a ɲɛfɛ. "A tɛmɛna sogo nin na ka taga ɲɛfɛ, ka kɔlɔn saba ye. Fɔlɔ jalen bɛ, ji foyi t'a la. A filanan ji to kɔnɔ. A sabanan ji b'o kɔnɔ. A tɛmɛn'o la ka taga se cɛkɔrɔnin dɔ ma. Cɛ nin kɔrɔla kosɛbɛ. [...]

Fichier Daba

Nsiirin, nsiirin.

Nsiirin	,	nsiirin	.
nsíirin		nsíirin	
n		n	
conte		conte	

N y'a bila den dɔ le kan.

N	y'	a	bila	den	dɔ	le	kan	.
ń	y'	à	bìla	dén	dɔ́	le	kàn	
pers	pm	pers	v	n	dtm	prt	pp	
1SG	PFV.TR	3SG	mettre	enfant	certain	FOC	sur	

Fichier vertical

Token	Lemma	PdD	Glose	Compos.	Original	Tonal
nsiiri	nsiiri	n	conte		NSIIRI	nsiiri
naaninan	naaninan	ORD|adj	quatrième	naani	NAANINAN	náaninan
dinye	dunya|jinye |dunu-nye|dinye |jyen|diyen	n	monde		Diɲɛ	díɲɛ
yaalala	yaalala	AG.PRM|n		yaala	Yaalala	yáalala
nsiirin	nsiiri|nsiirin	n	conte		Nsiirin	nsíirin
,	,	c	,	,	,	,
nsiirin	nsiiri|nsiirin	n	conte		nsiirin	nsíirin
.	.	c	.	.	.	.
n	n	pers	1SG		N	ń
y'	ye|y'	pm	PFV.TR		y'	y'
a	a	pers	3SG		a	à
bila	bil'|bla|bila	v	mettre		bila	bìla
den	den	n	enfant		den	dén
do	do	dtm	certain		dɔ	dɔ́
le	le	prt	FOC		le	le
kan	kan	pp	sur		kan	kàn
.	.	c	.	.	.	.

Volumétrie des ressources

- Corpus

Corpus	Balises	Ponctuations	Formes (distinctes)
Brut	412K	383K	2 321K (68K)
Désambiguïsé	104K	71K	426K (19K)

Volumétrie des ressources

- Corpus

Corpus	Balises	Ponctuations	Formes (distinctes)
Brut	412K	383K	2 321K (68K)
Désambiguïsé	104K	71K	426K (19K)

- Dictionnaires (disponibles en ligne)

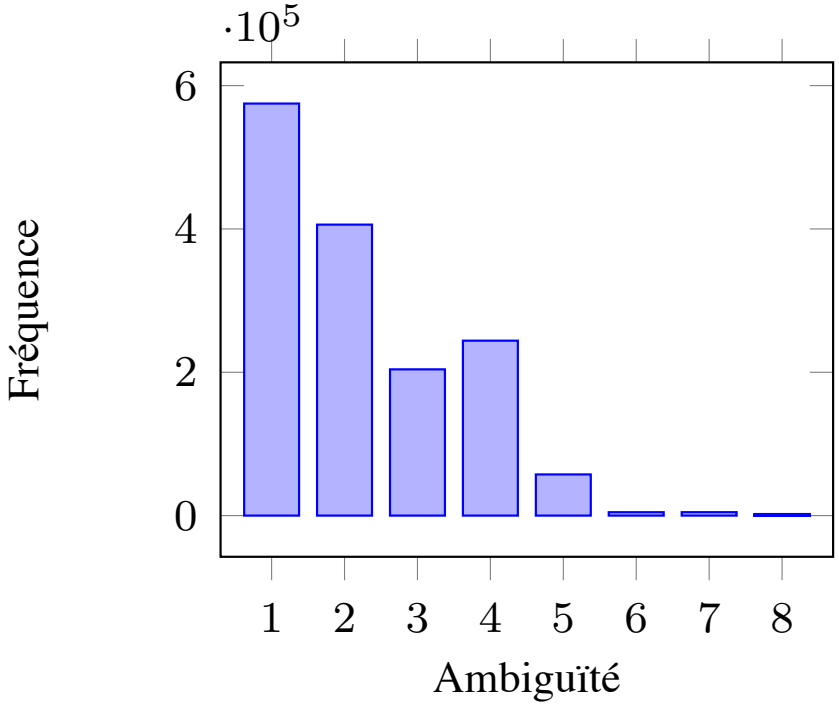
Dictionnaire	Description	Entrées	Ambiguïté
bamadaba	Dictionnaire principal	11K	1,137
enciclop	Notions encyclopédiques	29	1
jamuw	Noms claniques	375	1,001
togow	Prénoms	496	1
yorow	Toponymes	299	1

Statistiques sur les ambiguïtés

- Corpus échantillon (1,5M mots)
  - Traité avec Daba
  - 1,285M mots (215K ponctuations)
  - 72K (5,6%) parties du discours non renseignées
  - 288K (22%) non-ambiguës (selon Daba)

Statistiques sur les ambiguïtés

- Corpus échantillon (1,5M mots)
  - Traité avec Daba
  - 1,285M mots (215K ponctuations)
  - 72K (5,6%) parties du discours non renseignées
  - 288K (22%) non-ambiguës (selon Daba)



Ambiguïté	Fréquence
pers|intj	75K
n|v	72K
cop|pp|v|pm	57K
adv|conj|pm|v	53K
cop|pm	43K
v|n	38K
prt|conj|prn	33K
n.prop|pp|conj|intj	23K
conj|n|prep|v	23K
vq|v|adj|n|cop	22K

Caractérisation des ambiguïtés

- Ambiguïtés fréquentes

Mot	Ambiguïté	Fréquence
à	pers|intj	74K
yé	cop|pp|v|pm	58K
kà	adv|conj|pm|v	53K
ò	prt|conj|prn	34K

Caractérisation des ambiguïtés

- Ambiguïtés fréquentes

Mot	Ambiguïté	Fréquence
à	pers|intj	74K
yé	cop|pp|v|pm	58K
kà	adv|conj|pm|v	53K
ò	prt|conj|prn	34K

- Par mots distincts

Ambiguïté	Fréquence
PL|n	1304
n|v	1189
v|n	1039
ptcp|PTCP.RES	588
PFV.INTR|v	584
NMLZ|n	568
n.prop|dtm|ptcp|prn|adj|PL|n	497
n.prop|n	483

Jeu d'étiquettes sur le corpus désambiguisé

Code	Partie du discours	Quantité
n	nom	82K
c	ponctuation	66K
pers	pronom personnel	54K
v	verbe	51K
pm	marque prédicative	41K
pp	postposition	34K
conj	conjonction	21K
cop	copule	18K
n.prop	nom propre	12K
dtm	déterminatif	12K
prn	pronom (non-personnel)	10K
prt	particule	10K

num	numératif	6K
adj	adjectif	4K
ptcp	participe	4K
intj	interjection	2K
adv	adverbe	2K
vq	verbe qualitatif	1K
onomat	onomatopée	102
adv.p	adverbe préverbal	26
conv.n	converbe nu	24
mrph	morphème	13

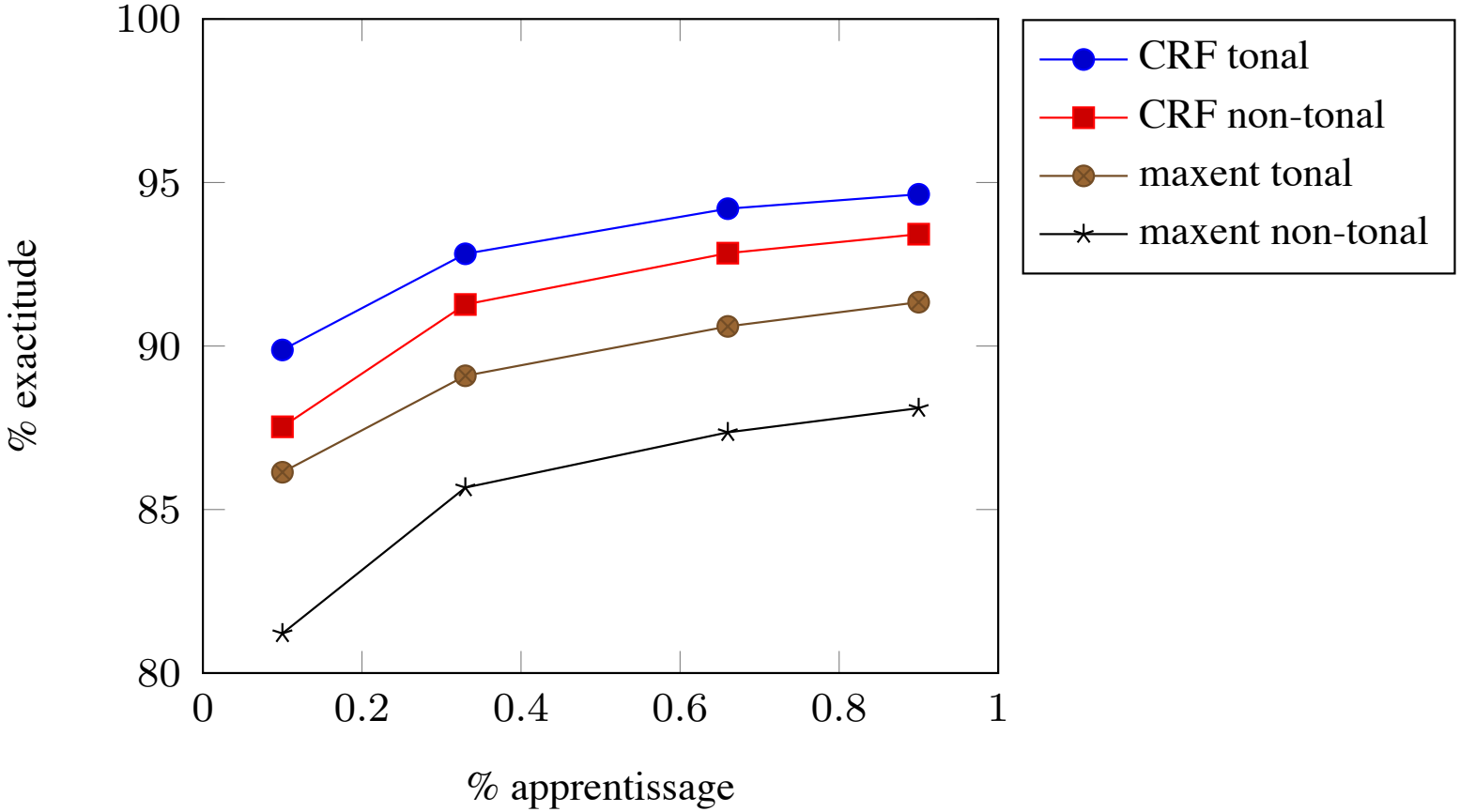
⇒ Jeu d'étiquettes relativement standard

Apprentissage : algorithme et données

- Méthodologie
  - Utilisation de Wapiti (Lavergne, 2010)
  - Validation croisée à 10 plis

Apprentissage : algorithme et données

- Méthodologie
  - Utilisation de Wapiti (Lavergne, 2010)
  - Validation croisée à 10 plis



Traits pour l'apprentissage

- Ajouts de traits pour
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Une version transformé du mot par un expression régulière, ANP
  - Une version non-tonalisée du mot
  - La taille du mot
  - Les étiquettes possibles dans les dictionnaires

Traits pour l'apprentissage

- Ajouts de traits pour
  - Préfixe de 2 et 3 caractères
  - Suffixe de 2 et 3 caractères
  - Une version transformé du mot par un expression régulière, ANP
  - Une version non-tonalisée du mot
  - La taille du mot
  - Les étiquettes possibles dans les dictionnaires

Score	Base	Non-tonal	Préf.	Suf.	Dico	ANP	Taille	Tous
Mot U	86.14	86.36	89.09	89.46	89.87	89.41	89.14	90.65
Phrase U	18.70	19.09	24.38	25.60	25.62	25.18	23.82	29.04
Mot B	85.94	91.27	91.69	91.91	91.50	86.84	88.37	94.22
Phrase B	13.14	35.35	36.39	37.51	33.94	15.20	21.57	47.90

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Comparaison avec TreeTagger

- Tests TreeTagger
  - Entraînement : 90% du corpus, validation croisée
  - Pas de ressource additionnelle
  - Configuration par défaut

Outil	**Score**
Baseline	22%
Majorité	82,06%
TreeTagger	93,50
Wapiti	94.22
Majorité (non-tonal)	79,68%

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)
  - ⇒ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

Utilisation des tons

- Caractéristiques
  - Trois marques tonales : ̀, ́, ̌ (caron, hatchek)
  - ⇒ Change le sens du mot
  - Exemples :
    - bá = maman / bà = chèvre
    - tùgu = bras / túgu = fermer
    - tà = prendre, porter / tá = feu, propriété
- Les tons sont peu souvent marqués à l'écrit
- Essentiellement en 1 ère syllabe

⇒ La présence de tons aide pour la morpho-syntaxe

⇒ Détecter automatiquement les tons ?

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

$\Rightarrow$ 1 518 tonalisations ambiguës, en moyenne 2,26 tonalisations

Statistiques sur les tons

- Méthodologie
  - Sélection du corpus en version tonale (traitée)
  - Suppression des tons et comparaison des versions non-tonales
  - Tests de « re-tonalisation »

$\Rightarrow$ 17 335 formes non-tonales, en moyenne 1.11 tonalisations

$\Rightarrow$ 1 518 tonalisations ambiguës, en moyenne 2,26 tonalisations

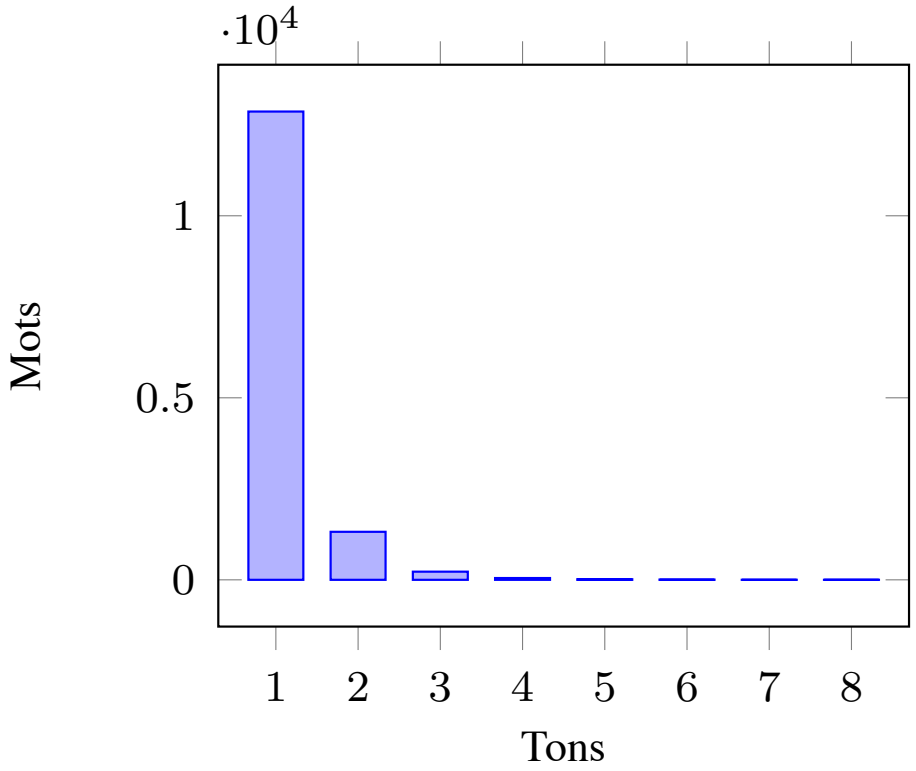
$\Rightarrow$ Baseline, tonalisation la plus fréquente : 59% (dont ponctuations)

Répartitions des tons

- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots

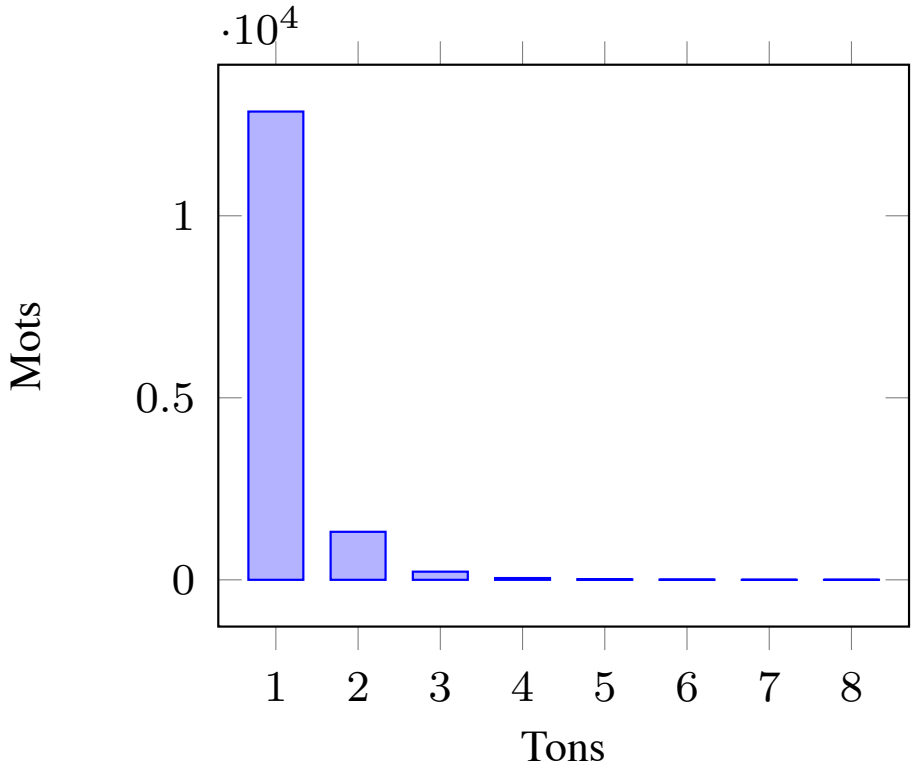
Répartitions des tons

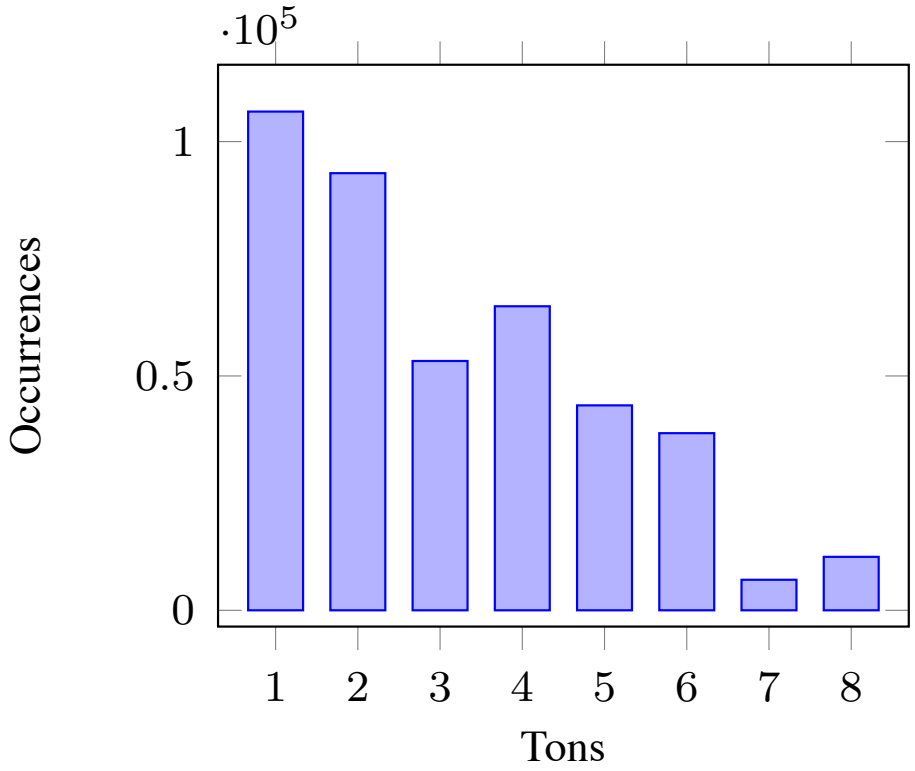
- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots



Répartitions des tons

- Méthodologie
  - Répartition du nombre de tonalisations dans le corpus
  - Comparaison : mots / occurrences de mots





Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Entropie des tons

- Méthodologie
  - Probabilités de tonalisations
  - Calcul d'entropie par mot

Ent.	Tonalisations	Traduction
3.20	táatúmà : 1.0 , tàatúmá : 1.0 , táatúmá : 1.0 , táatùma : 9.0 , táatúma : 1.0 , tàatúmà : 1.0 , tàatúma : 1.0 , tàatùmà : 1.0 , tàatùmá : 1.0 , táatùmá : 1.0 , táatùmà : 2.0 , tàatuma : 1.0 , tàatùma : 1.0 , táatuma : 1.0	Départ (?)
2.50	bámànankɛ : 10.0 , bámànakɛ̀ : 16.0 , bámànankɛ̀ : 5.0 , bámànánkɛ : 6.0 , cɛ̀ : 16.0 , bámànanke : 1.0 , bámànan : 16.0 , bámanankɛ : 1.0	Bambara
2.25	súurun : 1.0 , súuru : 1.0 , sùruntu : 2.0 , súruntu : 1.0 , sùrundu : 1.0	Verser
2.02	cíyɛn : 9.0 , tìɲɛ : 11.0 , tíɲɛ : 2.0 , ciyɛ́n : 2.0 , cìyɛn : 11.0	Vérité
2.00	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ : 1.0	Regard
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Approcher
1.99	ɲɛ̀nen : 1.0 , ɲɛ́nen : 1.0 , ɲànen : 1.0 , ɲɛ̀ : ? 1.0	?
1.99	ɲɔ́gɔri : 7.0 , nwàri : 1.0 , ɲɔ̀gɔri : 1.0 , ɲɔ́ɔri : 5.0 , ɲúari : 5.0	Salir
1.99	lé : 1.0 , lè : 4.0 , Ìè : 1.0 , le : 7.0 , dè : 4.0	(clan)
1.95	tàamasyɛnw : 2.0 , tàamashyɛn : 2.0 , táamashyɛnw : 1.0 , tàamashyɛnw : 2.0	Indiquer

Tonalisation automatique

- Stage M1 « Désambiguïsation lexicale des corpus bambara » (E. Mboning, INALCO)
- En moyenne deux tonalisations possible par mot non-tonalisé
- Apprentissage automatique (NLTK)
- Division train / test : 80% / 20%
- Essai de plusieurs algorithmes
  - CRF (peu de features) : 73%
  - Séquetiel (bigrammes, trigrammes, etc.) : 80%

Conclusions

- Corpus du bambara
  - Initiative déjà ancienne
  - Corpus plutôt TAL-compatible
- Analyse en parties du discours
  - Assez bonnes performances
  - Problématique de tonalisation
- Tonalisation
  - Utile à la morpho-syntaxe
  - Probablement utile pour les lemmes et gloses

Plan

1. La désambiguisation lexicale : des formes aux sens

2. Reconnaissance et liaison des entités nommées

3. Enrichissement lexical d'un corpus bambara

4. Quelques travaux préliminaires sur l'arabe

5. Conclusions

Translittérations de l'arabe

- Problématique qui mêle
  - Des **langues**
  - La **parole**
  - Les **systèmes d'écritures**

Translittérations de l'arabe

- Problématique qui mêle
  - Des **langues**
  - La **parole**
  - Les **systèmes d'écritures**

$\Rightarrow$ Transformations **très** sensibles aux ambiguités / erreurs

Pour lire de l'arabe

- Stage M2 « Translittération de l'arabe : Calcul des énoncés sapientiels médiévaux multilingues » (M. Khlifi, INALCO)
- Projet Aliento
  - Énoncés « spaientiels »
  - **Langues** : latin, arabe, hébreu, espagnol, catalan
  - **Époque** : XII ème siècle

  ⇒ Connaître les « rythmes » des proverbes arabes sans les lire ?

⇒ Translittération : langue et écriture **arabe** vers l'écriture **latine**

- Logiciel MADAMIRA (Columbia)
- Nécessité de « voyeller » l'arabe
- Levenshtein : 20% (sans voyelles) à 50% (avec voyelles)

⇒ Ajout de voyelles comme **désambiguisation contextuelle**

Exemple de translittération

- Texte sans voyelles :

نقوش فصوص خواتيم الفلاسفة

Nqwsh Fsws Khwatym Al-Flasft

- Texte avec voyelles

نُقُوشٍ فَصُوصِ خَواتِيمِ الفَلاسِفَةِ

Nuqoushin Fasousi Khawateimi Al-Falasifati

Désambiguiser pour traiter de l'arabizi

- Stage M2 « Translittération et normalisation de la langue arabe pour l'analyse de sentiments dans les médias sociaux » (L. Feredj, Synthesio)

Désambiguiser pour traiter de l'arabizi

- Stage M2 « Translittération et normalisation de la langue arabe pour l'analyse de sentiments dans les médias sociaux » (L. Feredj, Synthesio)
- Langue **arabe**, mais écriture **latine**…

⇒ Arabizi, franco-arab, arabish, etc.

Désambiguiser pour traiter de l'arabizi

- Stage M2 « Translittération et normalisation de la langue arabe pour l'analyse de sentiments dans les médias sociaux » (L. Feredj, Synthesio)
- Langue **arabe**, mais écriture **latine**…

⇒ Arabizi, franco-arab, arabish, etc.

  - Objectif : traiter des textes en arabizi pas des systèmes TAL

  ⇒ Translittération « inverse » ?!

Désambiguiser pour traiter de l'arabizi

- Stage M2 « Translittération et normalisation de la langue arabe pour l'analyse de sentiments dans les médias sociaux » (L. Feredj, Synthesio)
- Langue **arabe**, mais écriture **latine**…

⇒ Arabizi, franco-arab, arabish, etc.

- Objectif : traiter des textes en arabizi pas des systèmes TAL

⇒ Translittération « inverse » ?!

- Peu d'ambiguités (voyelles du français)
- Système à base de règles
  - Morphologiques : préfixes, suffixes, duplications, etc.
  - Priorité des règles
  - Validation par méthode lexicale (dictionnaire de mots)

⇒ Performances (Levenshtein) : 72% (mot, dont 90% $\leq$ 1 erreur)

Exemple de translittération

3ala	على
3adi	عادي
3lik	عليك
3lih	عليه
7ata	حتى
7aja	حاجة
7sal	حصل
9rit	قريت
9fas	قفص

Plan

1. La désambiguisation lexicale : des formes aux sens

2. Reconnaissance et liaison des entités nommées

3. Enrichissement lexical d'un corpus bambara

4. Quelques travaux préliminaires sur l'arabe

5. Conclusions

Conclusions

- La désambiguisation comme...
  - Processus d'**analyse linguistique** (POS)
  - Traitement et **enrichissement** de corpus
  - **Liaison** entre textes et bases de connaissances

Conclusions

- La désambiguisation comme...
  - Processus d'**analyse linguistique** (POS)
  - Traitement et **enrichissement** de corpus
  - **Liaison** entre textes et bases de connaissances

$\Rightarrow$ Mécanisme pour la compréhension du langage

Conclusions

- La désambiguisation comme...
  - Processus d'**analyse linguistique** (POS)
  - Traitement et **enrichissement** de corpus
  - **Liaison** entre textes et bases de connaissances

$\Rightarrow$ Mécanisme pour la compréhension du langage

- Exploitation de
  - **Bases lexicales** (problèmes de **couverture**)
  - Modèles contextuels (déjà au niveau du document ...)

Quelques perspectives

- Beaucoup de travaux **génériques**, mais
  - **Domaines** de spécialité
  - **Temporalité** / diachronie des textes
  - Diversité des **langues** et **dialectes**

Quelques perspectives

- Beaucoup de travaux **génériques**, mais
  - **Domaines** de spécialité
  - **Temporalité** / diachronie des textes
  - Diversité des **langues** et **dialectes**
- Apprentissage automatique peu interactif
  - Liens entre analyse **syntaxique** et entités
  - Exploitation d'indices **coréférentiels**
  - Problème des **signaux faibles**

Quelques perspectives

- Beaucoup de travaux **génériques**, mais
  - **Domaines** de spécialité
  - **Temporalité** / diachronie des textes
  - Diversité des **langues** et **dialectes**
- Apprentissage automatique peu interactif
  - Liens entre analyse **syntaxique** et entités
  - Exploitation d'indices **coréférentiels**
  - Problème des **signaux faibles**
- Avancer dans la **modélisation de la désambiguisation**
  - Quels traits pour la liaison ?
  - Mélange d'indices textuels et issus des connaissances
  - Approches moins **gourmandes** en ressources