

**Ponctuation et syntaxe**  
dans la langue française médiévale  
Étude d'un corpus de chartes originales  
écrites à Liège entre 1236 et 1291  
**L'envers des tapisseries**

Nicolas Mazziotta

21 décembre 2007, exposé de soutenance

# Plan

## 1 Rappel du plan de la thèse

# Plan

- 1 Rappel du plan de la thèse
- 2 Techniques d'annotation

# Plan

- 1 Rappel du plan de la thèse
- 2 Techniques d'annotation
- 3 Prolongements

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

- 1** Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2** Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
  
- 3** Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

# Question de départ et besoin de modélisation

## Situation de départ

# Question de départ et besoin de modélisation

## Situation de départ

- Pari : la syntaxe explique les signes de ponctuation



# Question de départ et besoin de modélisation

## Situation de départ

- Pari : la syntaxe explique les signes de ponctuation
- Servons-nous du corpus de *Khartês*

# Question de départ et besoin de modélisation

## Situation de départ

- Pari : la syntaxe explique les signes de ponctuation
- Servons-nous du corpus de *Khartês*

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la **punctuation** originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la **punctuation** ?

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la **syntaxe** dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la ponctuation ?
- Qu'est-ce que la **syntaxe** ?

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la ponctuation ?
- Qu'est-ce que la syntaxe ?

Répondre à ces questions avant toute chose



# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la ponctuation ?
- Qu'est-ce que la syntaxe ?

Répondre à ces questions avant toute chose  
( Question du corpus résolue de manière pratique )

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la ponctuation ?
- Qu'est-ce que la syntaxe ?

Répondre à ces questions avant toute chose

↳ Besoin (théorique) de **définir les concepts**

# Question de départ et besoin de modélisation

## Guide pour toute la recherche

Comment, d'après ce qu'on peut observer dans les chartes écrites en français à Liège avant 1292, la ponctuation originale interagit-elle avec la syntaxe dans la langue française médiévale ?

## Implique des sous-questions

- Quel corpus ?
- Qu'est-ce que la ponctuation ?
- Qu'est-ce que la syntaxe ?

Répondre à ces questions avant toute chose

➔ Besoin (théorique) de **définir les concepts**

Objet de la première partie

## 1 Rappel du plan de la thèse

- Question de départ et besoin de modélisation
- **Analyse des données : statistiques**
- Conclusions

## 2 Techniques d'annotation

- Un travail caché
- Langage de base : XML (présentation)
- Problématique de l'annotation
- Interface

## 3 Prolongements

- Élargissement du corpus, projet de recherche
- Partage des techniques

# Analyse des données : statistiques

## Démarche

# Analyse des données : statistiques

## Démarche

Chiffrer les liens entre les structures et les signes de ponctuation

## Travail en trois étapes

# Analyse des données : statistiques

## Démarche

Chiffrer les liens entre les structures et les **ponctogrammes**

## Travail en trois étapes

- 1 Repérer les tendances générales

# Analyse des données : statistiques

## Démarche

Chiffrer les liens entre les structures et les **ponctogrammes**  
Définir les individus : structures et ponctogrammes

## Travail en trois étapes

- 1 Repérer les tendances générales
- 2 Focaliser l'analyse sur ces tendances



# Analyse des données : statistiques

## Démarche

Chiffrer les liens entre les structures et les **punctogrammes**

Définir les individus : structures et punctogrammes

↳ Nécessaire simplification

## Travail en trois étapes

- 1 Repérer les tendances générales
- 2 Focaliser l'analyse sur ces tendances
- 3 Analyser la forme des punctogrammes

## 1 Rappel du plan de la thèse

- Question de départ et besoin de modélisation
- Analyse des données : statistiques
- Conclusions

## 2 Techniques d'annotation

- Un travail caché
- Langage de base : XML (présentation)
- Problématique de l'annotation
- Interface

## 3 Prolongements

- Élargissement du corpus, projet de recherche
- Partage des techniques

# Conclusions

## Tendances repérées

# Conclusions

## Tendances repérées

- Cohérence et « règles »

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

## Subjectivité

Vision personnelle

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

## Subjectivité

Vision personnelle

↳ Nos définitions marquent l'ensemble



# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

## Subjectivité

Vision personnelle

➔ Nos définitions marquent l'ensemble

## Portes ouvertes

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

## Subjectivité

Vision personnelle

➔ Nos définitions marquent l'ensemble

## Portes ouvertes

Autres matériaux

# Conclusions

## Tendances repérées

- Cohérence et « règles »
- Importance des **absences**
- **Remise en question** du modèle syntaxique

## Subjectivité

Vision personnelle

➔ Nos définitions marquent l'ensemble

## Portes ouvertes

Autres matériaux

Exploitation de chaque modèle séparément

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

# Un travail caché

Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

# Un travail caché

Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche



# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ construction de ces modèles

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ construction de ces modèles
  - ↳ et leur application au corpus

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ construction de ces modèles
  - ↳ et leur application au corpus
- Analyses statistiques articulées

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ construction de ces modèles
  - ↳ et leur application au corpus
- Analyses statistiques articulées
  - ↳ choix des tests

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ construction de ces modèles
  - ↳ et leur application au corpus
- Analyses statistiques articulées
  - ↳ choix des tests

Ce qui nous intéresse ici

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ ~~construction de ces modèles~~
  - ↳ et leur application au corpus
- Analyses statistiques articulées
  - ↳ choix des tests

Ce qui nous intéresse ici

# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ ~~construction de ces modèles~~
  - ↳ et leur application au corpus
- Analyses statistiques articulées
  - ↳ ~~choix des tests~~

Ce qui nous intéresse ici



# Un travail caché

## Focalisation de la thèse sur la dimension scientifique

Mise en évidence du **résultat**

↳ la progression n'est pas montrée

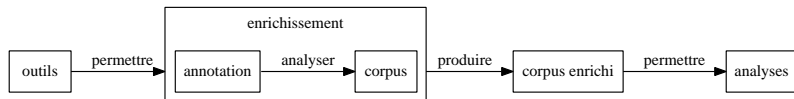
Cette progression soutient toute la recherche

- Modèles construits argumentés
  - ↳ ~~construction de ces modèles~~
  - ↳ et leur **application au corpus**
- Analyses statistiques articulées
  - ↳ ~~choix des tests~~

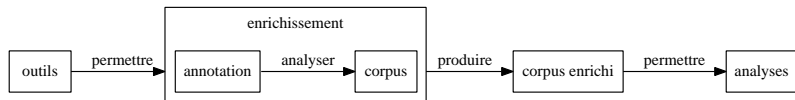
Ce qui nous intéresse ici

Comment avons-nous procédé... **concrètement** ?

# Schématiquement



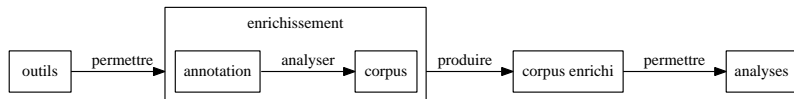
# Schématiquement



## Une série de choix

- Le corpus est **électronique** (déterminé au niveau du projet)
- Nous voulons employer les **statistiques**

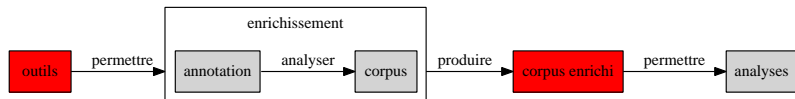
# Schématiquement



## ↳ Besoin d'annotations électroniques

- s'intègrent au corpus
- peuvent être extraites automatiquement

# Schématiquement



## Ce qui nous intéresse ici

- Quels outils/techniques permettent d'enrichir le corpus ?
- À quoi ressemble le corpus (avant et après) ?

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - **Langage de base : XML (présentation)**
  - Problématique de l'annotation
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

Texte et balisage

```
<balisage>texte</balisage>
```

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

Texte et balisage

```
<balisage>texte</balisage>
```

Balise de début (nom : *balisage*)



# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

Texte et balisage

```
<balisage>texte</balisage>
```

Balise de fin (même nom)

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>
```

```
<balisage attribut="valeur">texte</balisage>
```

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>
```

```
<balisage attribut="valeur">texte</balisage>
```

Attribut (paire *clé*="valeur")

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

Texte

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

## Balisage

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

*h1* indique les limites du titre

# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

*p* marque le paragraphe



# Langage de base : XML (présentation)

Emploi de la norme XML (portée internationale, consensuelle, moderne)

## Texte et balisage

```
<balisage>texte</balisage>  
<balisage attribut="valeur">texte</balisage>
```

## Application : le XHTML (pour les pages web)

```
<h1>Introduction</h1>  
<p>Texte de l'introduction</p>
```

Principes utiles pour **indiquer la structure** des textes

# Langage de base : XML (présentation)

## Hiérarchie du balisage

Récurtivité : `<balise><balise>Texte</balise></balise>`

# Langage de base : XML (présentation)

## Hiérarchie du balisage

Récurtivité : `<balise><balise>Texte</balise></balise>`

**Pas de limite** à l'emboîtement

# Langage de base : XML (présentation)

## Hiérarchie du balisage

Réursivité : `<balise><balise>Texte</balise></balise>`

**Pas de limite** à l'emboîtement

Mais le balisage ne peut se chevaucher

# Langage de base : XML (présentation)

## Hiérarchie du balisage

Récurtivité : `<balise><balise>Texte</balise></balise>`

**Pas de limite** à l'emboîtement

Mais le balisage ne peut se chevaucher

## Application : XHTML

```
<div>
  Cadre
  <div> 1er sous-cadre </div>
  <div> 2e sous-cadre </div>
</div>
```

# Langage de base : XML (présentation)

## Hiérarchie du balisage

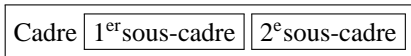
Récurtivité : `<balise><balise>Texte</balise></balise>`

**Pas de limite** à l'emboîtement

Mais le balisage ne peut se chevaucher

## Application : XHTML

```
<div>  
  Cadre  
  <div> 1er sous-cadre </div>  
  <div> 2e sous-cadre </div>  
</div>
```



- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - **Problématique de l'annotation**
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

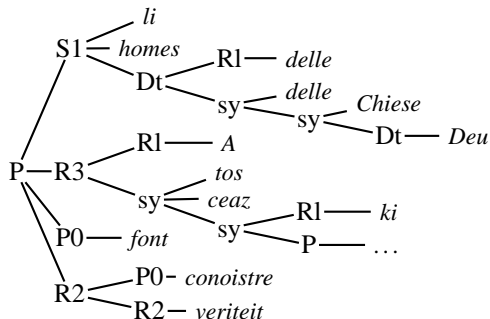
# Problématique de l'annotation

A tos ceaz ki ces presens letres veront *et* oront, li homes delle Chiese Deu font conoistre veri-<sup>[2]</sup>-teit.



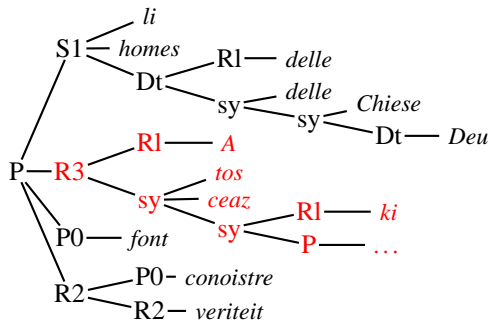
# Problématique de l'annotation

A tos ceaz ki ces presens lettres veront *et* oront, li homes delle Chiese Deu font conoistre veri-<sup>[2]</sup>-teit.



# Problématique de l'annotation

*A tos ceaz ki ces presens letres veront et oront*, li homes delle Chiese Deu font conoistre veri-<sup>[2]</sup>-teit.



## XML

```
<R3>
  <R1>A</R1>
  <Sy>tos ceaz
  <Sy>
    <R1>ki</R1>
    <P>...</P>
  </Sy>
</R3>
```

# Problématique de l'annotation

Que se passe-t-il si on procède ainsi ?

Imaginons une analyse diplomatique concurrente

A tos ceaz ki ...

li homes delle Chiese Deu

font conoistre veriteit

# Problématique de l'annotation

Que se passe-t-il si on procède ainsi ?

Imaginons une analyse diplomatique concurrente

```
<protocole>
  <adresse>
    A tos ceaz ki ...
  </adresse>
  <suscription>
    li homes delle Chiese Deu
  </suscription>
  <salut>
    font conoistre veriteit
  </salut>
</protocole>
```

# Problématique de l'annotation

Que se passe-t-il si on procède ainsi ?

Imaginons une analyse diplomatique concurrente

```
<R3>A tos ceaz ki ...</R3>
```

```
<S1>li homes delle Chiese Deu</S1>
```

```
<P0>font </P0><R2>conoistre veriteit</R2>
```

# Problématique de l'annotation

Que se passe-t-il si on procède ainsi ?

Imaginons une analyse diplomatique concurrente

```
<protocole>  
  <adresse>  
    <R3>A tos ceaz ki ...</R3>  
  </adresse>  
  <suscription>  
    <S1>li homes delle Chiese Deu</S1>  
  </suscription>  
  <salut>  
    <P0>font </P0><R2>conoistre veriteit</R2>  
  </salut>  
</protocole>
```

# Problématique de l'annotation

## Que se passe-t-il si on procède ainsi ?

### Imaginons une analyse diplomatique concurrente

```
<protocole>  
  <adresse>  
    <R3>A tos ceaz ki ...</R3>  
  </adresse>  
  <suscription>  
    <S1>li homes delle Chiese Deu</S1>  
  </suscription>  
  <salut>  
    <P0>font </P0><R2>conoistre veriteit</R2>  
  </salut>  
</protocole>
```

## Ce n'est ni convivial ni courtois

- On ne peut apposer d'annotations nouvelles sans **interférer** ;
- enrichir d'un point de vue **encombre** les autres.

# Problématique de l'annotation

Danger pour l'intégrité des données



# Problématique de l'annotation

## Danger pour l'intégrité des données

### Présence conjointe

- de l'édition, **qu'il ne faut pas toucher** ;
- de l'analyse, **qu'il faut construire**

# Problématique de l'annotation

## Danger pour l'intégrité des données

### Présence conjointe

- de l'édition, **qu'il ne faut pas toucher** ;
- de l'analyse, **qu'il faut construire**

↳ Risque d'erreur croissant avec l'annotation

# Problématique de l'annotation

## Danger pour l'intégrité des données

### Présence conjointe

- de l'édition, **qu'il ne faut pas toucher** ;
- de l'analyse, **qu'il faut construire**

↳ Risque d'erreur croissant avec l'annotation

## Inadéquation d'un point de vue épistémologique

Les annotations sont une **surcouché** par rapport à l'édition.

# Problématique de l'annotation

## Danger pour l'intégrité des données

### Présence conjointe

- de l'édition, **qu'il ne faut pas toucher** ;
- de l'analyse, **qu'il faut construire**

↳ Risque d'erreur croissant avec l'annotation

## Inadéquation d'un point de vue épistémologique

Les annotations sont une **surcouché** par rapport à l'édition.

↳ Nécessaire **disjonction** édition/annotations

# Problématique de l'annotation

## Danger pour l'intégrité des données

### Présence conjointe

- de l'édition, **qu'il ne faut pas toucher** ;
- de l'analyse, **qu'il faut construire**

↳ Risque d'erreur croissant avec l'annotation

## Inadéquation d'un point de vue épistémologique

Les annotations sont une **surcroupe** par rapport à l'édition.

↳ Nécessaire **disjonction** édition/annotations

Comment faire ?

# Problématique de l'annotation

Le principe est simple

# Problématique de l'annotation

## Le principe est simple

- chaque mot a une **identité**, chaque occurrence est **unique** ;

# Problématique de l'annotation

## Le principe est simple

- chaque mot a une **identité**, chaque occurrence est **unique** ;
- il suffit de dire à l'ordinateur :



# Problématique de l'annotation

## Le principe est simple

- chaque mot a une **identité**, chaque occurrence est **unique** ;
- il suffit de dire à l'ordinateur :
  - 1 comment donner une identité aux mots ;

# Problématique de l'annotation

## Le principe est simple

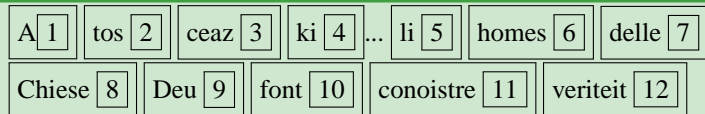
- chaque mot a une **identité**, chaque occurrence est **unique** ;
- il suffit de dire à l'ordinateur :
  - 1 comment donner une identité aux mots ;
  - 2 comment les retrouver.

# Problématique de l'annotation

## Le principe est simple

- chaque mot a une **identité**, chaque occurrence est **unique** ;
- il suffit de dire à l'ordinateur :
  - 1 comment donner une identité aux mots ;
  - 2 comment les retrouver.

## On pose alors des **étiquettes** sur les mots



# Problématique de l'annotation

XML peut être employé dans ce sens :

```
<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1 ">
  <c id="_1236">A</c>
</w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7">
  <c id="_1243">l</c>etres
</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

Balisateur comme identification

# Problématique de l'annotation

XML peut être employé dans ce sens :

```
<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1 ">
  <c id="_1236">A</c>
</w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7">
  <c id="_1243">l</c>etres
</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

Balisateur comme identification

Balises : limites des unités

# Problématique de l'annotation

XML peut être employé dans ce sens :

```
<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1 ">
  <c id="_1236">A</c>
</w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7">
  <c id="_1243">l</c>etres
</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

Balisateur comme identification

Attribut : identifiant

# Problématique de l'annotation

XML peut être employé dans ce sens :

```
<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1 ">
  <c id="_1236">A</c>
</w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7">
  <c id="_1243">l</c>etres
</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

Balisateur comme identification

Chaque mot a une **référence unique**

# Problématique de l'annotation

XML peut être employé dans ce sens :

```
<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1 ">
  <c id="_1236">A</c>
</w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7">
  <c id="_1243">l</c>etres
</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

Balisateur comme identification

Chaque mot a une **référence unique**

Principe

L'annotation ne consiste ici qu'à donner les **limites des unités** et à les **référer**



# Problématique de l'annotation

```
<link target="_1234"/>  
<link target="_381"/>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
...
```

## Balises comme pointeurs

# Problématique de l'annotation

```
<link target="_1234"/>  
<link target="_381"/>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
...
```

Balises comme pointeurs

Balise : unité vide

# Problématique de l'annotation

```
<link target="_1234"/>  
<link target="_381"/>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
...
```

Balises comme pointeurs

Attribut : référence

# Problématique de l'annotation

```
<link target="_1234"/>  
<link target="_381"/>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
...
```

Balises comme pointeurs

Pointe une **référence unique**

# Problématique de l'annotation

```
<link target="_1234"/>  
<link target="_381"/>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
...
```

Balises comme pointeurs

Pointe une **référence unique**

Principe

L'information est nulle : ce n'est qu'une **référence** aux données

# Problématique de l'annotation

## L'ordinateur peut relier les pointeurs aux références

```
<link target="_1234"/>          <l id="_1234"/>
<link target="_381"/>          <ponct id="_381">**</ponct>
<link target="_1" />           <w id="_1" ><c id="_1236">A</c></w>
<link target="_2" />           <w id="_2">tos</w>
<link target="_3" />           <w id="_3">ceaz</w>
<link target="_4" />           <w id="_4">ki</w>
<link target="_5" />           <w id="_5">ces</w>
<link target="_6" />           <w id="_6">pre$en$</w>
<link target="_7" />           <w id="_7"><c id="_1243">l</c>etres</w>
<link target="_8" />           <w id="_8">veront</w>
<link target="_9" />           <w id="_9">[et]</w>
<link target="_10" />          <w id="_10">oront</w>
...                               ...
```

# Problématique de l'annotation

## L'ordinateur peut relier les pointeurs aux références

```
<link target="_1234"/>
<link target="_381"/>
<link target="_1" />
<link target="_2" />
<link target="_3" />
<link target="_4" />
<link target="_5" />
<link target="_6" />
<link target="_7" />
<link target="_8" />
<link target="_9" />
<link target="_10" />
...

<l id="_1234"/>
<ponct id="_381">***</ponct>
<w id="_1" ><c id="_1236">A</c></w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7"><c id="_1243">l</c>etres</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```

# Problématique de l'annotation

## L'ordinateur peut relier les pointeurs aux références

```
<link target="_1234"/>
<link target="_381"/>
<link target="_1" />
<link target="_2" />
<link target="_3" />
<link target="_4" />
<link target="_5" />
<link target="_6" />
<link target="_7" />
<link target="_8" />
<link target="_9" />
<link target="_10" />
...

<l id="_1234"/>
<ponct id="_381">**</ponct>
<w id="_1" ><c id="_1236">A</c></w>
<w id="_2">tos</w>
<w id="_3">ceaz</w>
<w id="_4">ki</w>
<w id="_5">ces</w>
<w id="_6">pre$en$</w>
<w id="_7"><c id="_1243">l</c>etres</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
...
```



# Problématique de l'annotation

Les annotations sont disjointes du texte,  
mais aussi l'une de l'autre.

```
<adresse>  
<link target="_1" />  
<link target="_2" />  
<link target="_3" />  
<link target="_4" />  
<link target="_5" />  
<link target="_6" />  
<link target="_7" />  
<link target="_8" />  
<link target="_9" />  
<link target="_10" />  
</adresse>
```

```
<R3>  
<R1><link target="_1" /></R1>  
<Sy>  
  <link target="_2" />  
  <link target="_3" />  
<Sy>  
  <R1><link target="_4" /></R1>  
  <P>  
    <S1><link target="_4" /></S1>  
  <R2>  
    <link target="_5" />  
    <link target="_6" />  
    <link target="_7" />  
  </R2>  
<group> ... </group>  
<P>  
</Sy>  
</R3>
```

# Problématique de l'annotation

À l'aide des pointeurs, l'ordinateur peut  
recréer le lien si nécessaire.

```
<adresse>  
A  
tos  
ceaz  
ki  
ces  
presens  
letres  
veront  
et  
oront  
</adresse>
```

```
<R3>  
<R1>A</R1>  
<Sy>  
tos  
ceaz  
<Sy>  
<R1>ki</R1>  
<P>  
<S1>ki</S1>  
<R2>  
ces  
presens  
letres  
</R2>  
<group> ... </group>  
<P>  
</Sy>  
</Sy>  
</R3>
```

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
  
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
  
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

# Interface

The screenshot shows the Khedit application window. The main area displays a tree diagram of a sentence structure. The tree has a root node 'P' which branches into 'S1' and 'P0'. 'S1' branches into 'Jo' and 'Nicholes'. 'P0' branches into 'fai'. 'S1' also branches into 'syntagme'. 'R2' branches into 'P0' and 'R2'. The 'P0' under 'R2' branches into '\$a°uoir'. The 'R2' under 'R2' branches into 'Sb'. The 'Sb' under 'R2' branches into 'que'. The 'C5' branches into 'Sb'. The 'Sb' under 'C5' branches into 'cu[m]'. Below the tree is a table with columns: role, morpho, label, disjonctif, pc, note, aux, prop, rel.

role	morpho	label	disjonctif	pc	note	aux	prop	rel
P								pers
S1								
Jo								
Nicholes								
syntagme								
P0								
fai								
R2								inf
P0	inf	Vinf						
\$a°uoir								
R2								pers
Sb								_3092
que								
C5					cc			pers
Sb								_3096
cu[m]								

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
- 3 Prolongements
  - **Élargissement du corpus, projet de recherche**
  - Partage des techniques

# Élargissement du corpus, projet de recherche

Techniques et outils applicables à tout corpus XML

# Élargissement du corpus, projet de recherche

Techniques et outils applicables à tout corpus XML

↳ Élargissement possible



# Élargissement du corpus, projet de recherche

## Techniques et outils applicables à tout corpus XML

- ➔ Élargissement possible
- ➔ Projet de syntaxe « minimale » de l'afr.

- 1 Rappel du plan de la thèse
  - Question de départ et besoin de modélisation
  - Analyse des données : statistiques
  - Conclusions
- 2 Techniques d'annotation
  - Un travail caché
  - Langage de base : XML (présentation)
  - Problématique de l'annotation
  - Interface
- 3 Prolongements
  - Élargissement du corpus, projet de recherche
  - Partage des techniques

# Partage des techniques

## Traditionnellement

Les résultats de la recherche sont donnés à la communauté :

- articles, monographies
- corpus

# Partage des techniques

## Traditionnellement

Les résultats de la recherche sont donnés à la communauté :

- articles, monographies
- corpus

## Donner les outils

Nécessité de partager les techniques avec la communauté

# Partage des techniques

## Traditionnellement

Les résultats de la recherche sont donnés à la communauté :

- articles, monographies
- corpus

## Donner les outils

Nécessité de partager les techniques avec la communauté

➔ Post-doctorat à Lyon (BFM) dans cette optique

# MERCI