# Annoter facilement un corpus complexe : l'exemple de Pyrrha, interface de post-correction, et Pie, lemmatiseur et tagueur, pour l'ancien français

Ariane Pinche

# Annoter facilement un corpus complexe : l'exemple de Pyrrha, interface de post-correction, et Pie, lemmatiseur et tagueur morphosyntaxique, pour l'ancien français

**Ariane Pinche, Univ. Jean Moulin Lyon 3 et École nationale des chartes, UMR 5648 (CIHAM)**

Rencontres lyonnaises des jeunes chercheurs en linguistique historique, 6 juin 2019

# Introduction : la recherche et l'étiquetage linguistique des textes

- Chronophage, mais utile pour des études de corpus
    - Ex: les études stylométriques (lire Mellet, S. (2002). "La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ?" *Médiévales*, 21 (42), 13-26. https://doi.org/10.3406/medi.2002.1536)
- Une tâche qui occupe la recherche en sciences humaines (TAL) depuis l'émergence des outils numériques
- Développer une interface pour utiliser facilement les lemmatiseurs et corriger l'étiquetage automatique : Pyrrha

# 1. Lemmatisation d'un corpus

## 1.1 Présentation de Pie

Manjavacas, E., Kestemont, M., & Clérice, T. (2019). emanjavacas/pie v0.1.0. https://doi.org/10.5281/zenodo.1637878

- conçu à partir d'algorithmes d'intelligence artificielle
  - sans dictionnaire ;
  - sans règles prédéfinies ;
  - capable d'apprendre à partir de corpus déjà annotés ;

## 1.2 Les modèles d'annotation de Pyrrha

- Modèle "latin Lasla" : Deucalion Latin Lemmatizer. https://doi.org/10.5281/zenodo.2707476
- Modèle pour l'ancien français. https://doi.org/10.5281/zenodo.3237455
    - Lemmes issus du Tobler-Lommatzsch
    - Jeu d'étiquettes morphosyntaxiques issu du référentiel Cattex 2009 : Guillot, C., Prévost, S., & Lavrentiev, A. (2013). Manuel de référence du jeu Cattex09. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf

## Exemples d'annotation

| Form | Lemma | POS | Morph |
|---|---|---|---|
| conmence | comencier | VERcjg | MODE=ind TEMPS=pst PERS.=3 NOMB.=s |

*Table 1. Exemple d'annotation verbale*

| Form | Lemma | POS | Morph |
|---|---|---|---|
| signeur | seignor | NOMcom | NOMB.=s GENRE=m CAS=r |

*Table 2. Exemple d'annotation nominale*

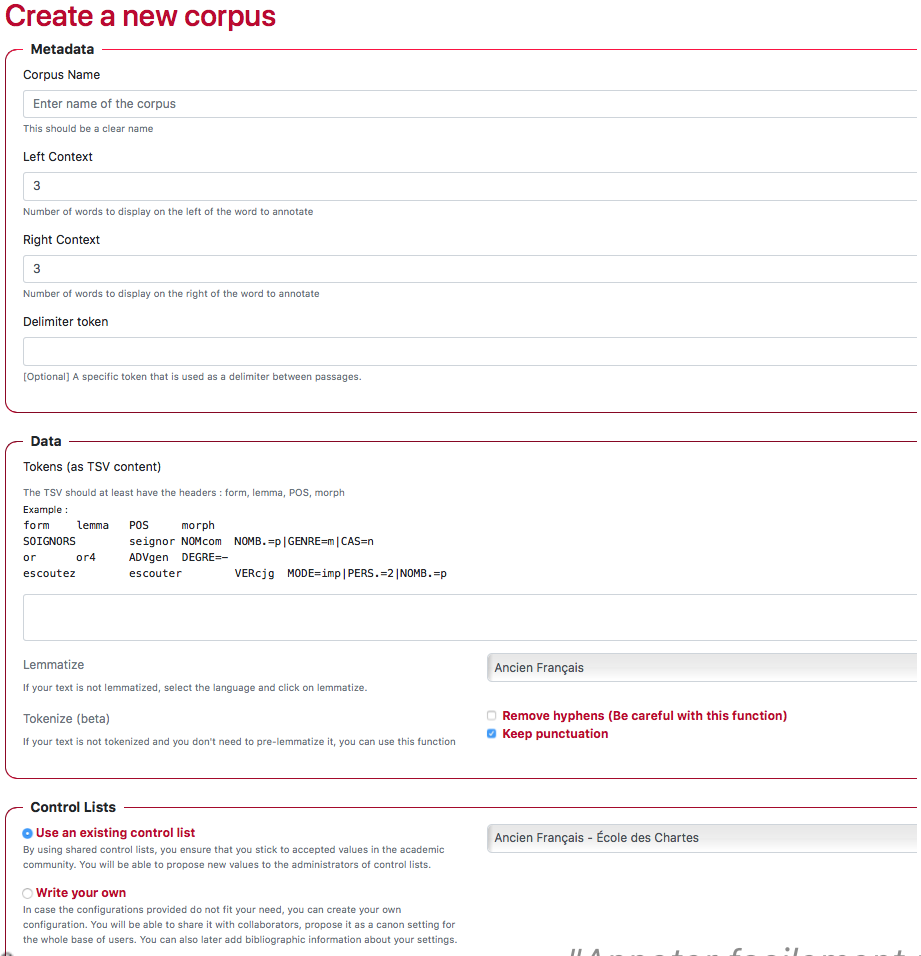
## 1.3 Lemmatiser son corpus à l'aide de Pyrrha

- https://dev.chartes.psl.eu/pyrrha/ pour apprendre à manipuler le corpus ;
- https://dh.chartes.psl.eu/pyrrha/ pour mettre en place un corpus pérenne.

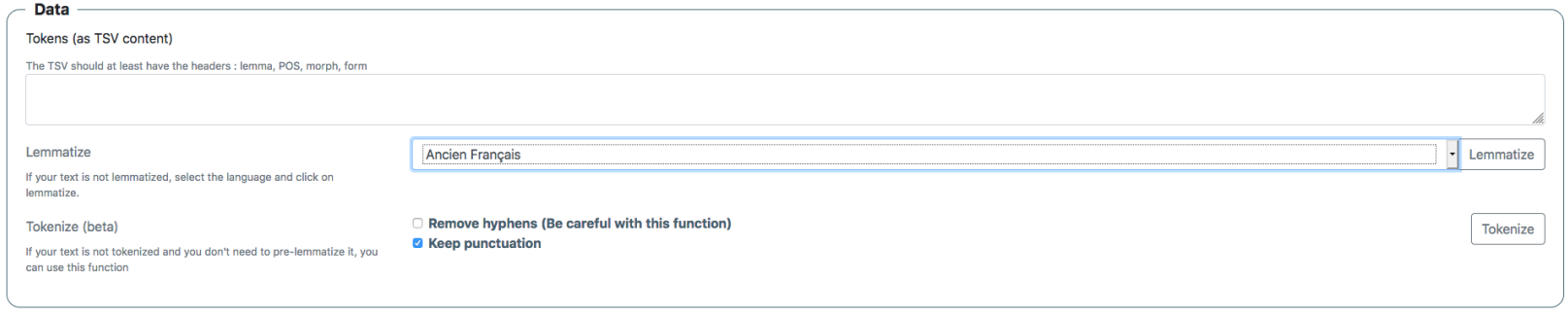
Condition requise : création d'un compte utilisateur.

# Création du corpus

- Nommer son corpus et définir la longueur de contexte à visualiser pour chaque mot du corpus dans l'interface de post-correction ;



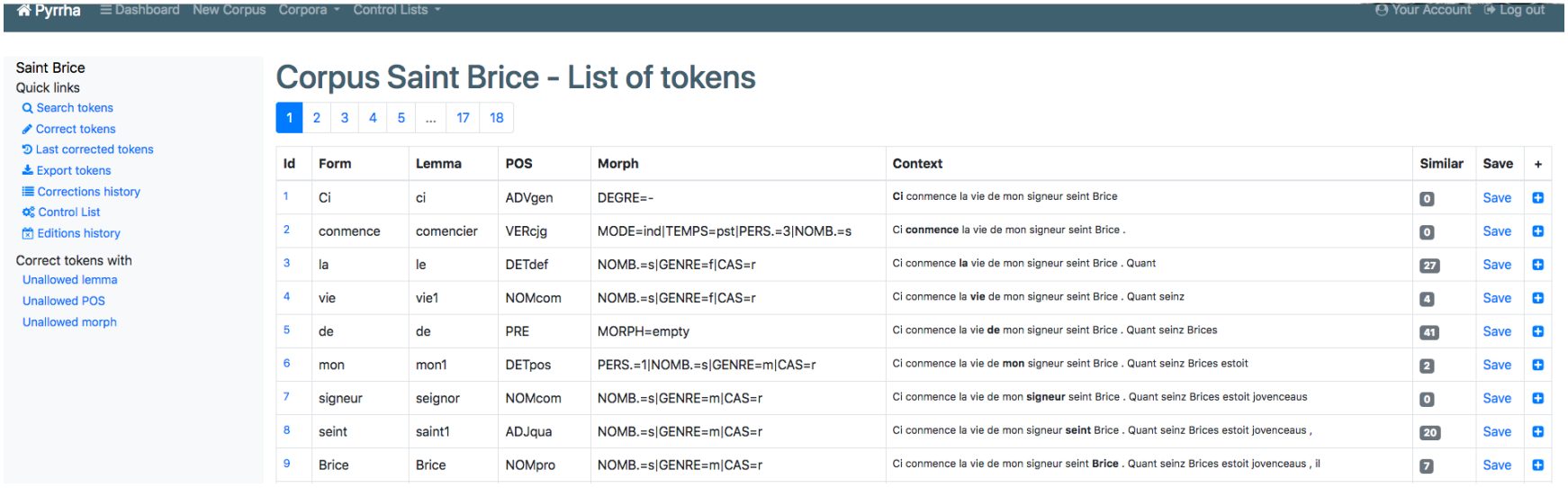
- Importer son corpus
    - copier-coller son texte brut au format texte dans la section data
    - copier-coller son texte déjà annoté au format TSV en faisant apparaître les champs suivants : *form, lemma, POS, morph*.
- Lemmatiser son corpus à l'aide de l'un des deux modèles proposés par l'interface




*Lemmatisation, capture d'écran, copyright © 2018 Clérice, Pilla, & Camps.*

- Choisir des listes de contrôle en choisissant parmi celles proposées ou bien en important la sienne
- Soumettre son corpus

# 2. Correction et export des données




| Id | Form | Lemma | POS | Morph | Context | Similar | Save | + |
|---|---|---|---|---|---|---|---|---|
| 1 | Ci | ci | ADVgen | DEGRE=- | **Ci** conmence la vie de mon signeur seint Brice | 0 | Save | |
| 2 | conmence | comencier | VERcjg | MODE=ind\|TEMPS=pst\|PERS.=3\|NOMB.=s | Ci **conmence** la vie de mon signeur seint Brice . | 0 | Save | |
| 3 | la | le | DETdef | NOMB.=s\|GENRE=f\|CAS=r | Ci conmence **la** vie de mon signeur seint Brice . Quant | 27 | Save | |
| 4 | vie | vie1 | NOMcom | NOMB.=s\|GENRE=f\|CAS=r | Ci conmence la **vie** de mon signeur seint Brice . Quant seinz | 4 | Save | |
| 5 | de | de | PRE | MORPH=empty | Ci conmence la vie **de** mon signeur seint Brice . Quant seinz Brices | 41 | Save | |
| 6 | mon | mon1 | DETpos | PERS.=1\|NOMB.=s\|GENRE=m\|CAS=r | Ci conmence la vie de **mon** signeur seint Brice . Quant seinz Brices estoit | 2 | Save | |
| 7 | signeur | seignor | NOMcom | NOMB.=s\|GENRE=m\|CAS=r | Ci conmence la vie de mon **signeur** seint Brice . Quant seinz Brices estoit jovenceaus | 0 | Save | |
| 8 | seint | saint1 | ADJqua | NOMB.=s\|GENRE=m\|CAS=r | Ci conmence la vie de mon signeur **seint** Brice . Quant seinz Brices estoit jovenceaus , | 20 | Save | |
| 9 | Brice | Brice | NOMpro | NOMB.=s\|GENRE=m\|CAS=r | Ci conmence la vie de mon signeur seint **Brice** . Quant seinz Brices estoit jovenceaus , il | 7 | Save | |

*Interface de correction, capture d'écran, copyright © 2018 Clérice, Pilla, & Camps*

## 2.1 Fonctionnalités de base : relecture et édition des corrections

L'accès à l'interface de post-correction se fait via l'onglet corpora pour sélectionner le texte à traiter. L'interface affiche un tableau avec neuf catégories différentes :

1. *Id* : un numéro est attribué à chaque token (mots et éléments de ponctuation) pour l'identifier ;
2. *Form* : terme tel qu'il apparaît dans le texte ;
3. *Lemma* : lemme attribué à chaque terme permettant ainsi de l'associer à une forme normalisée ;
4. *POS* : nature du mot ;

5. *Morph* : annotation morphosyntaxique ;
6. *Context* : le terme en gras accompagné de contexte textuel2;
7. *Similar* : nombre de termes dans une situation comparable ;
8. *Save* : sauvegarde les modifications opérées sur l'annotation ;
9. *+* : accès vers les options de modification du token : correction, suppression, ajout.

On peut intervenir directement sur les trois catégories suivantes :

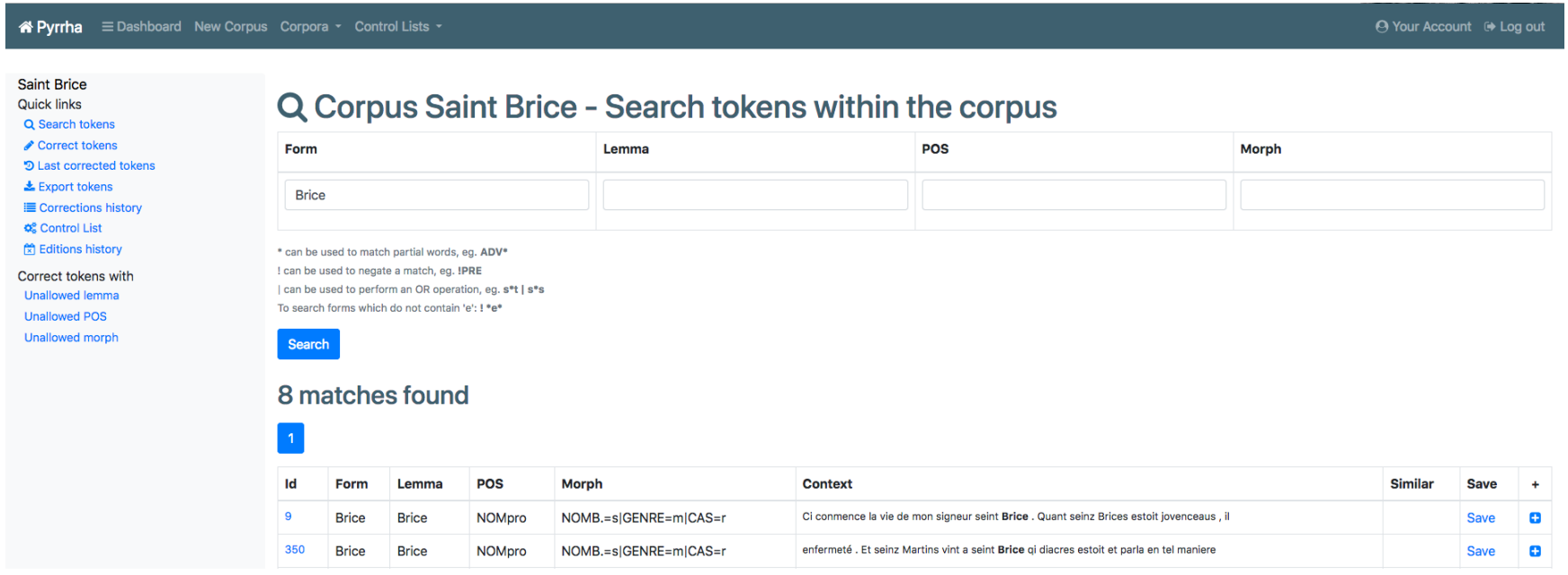
- lemme
- POS
- morph

**Attention** : Si jamais, l'une des catégories ainsi corrigées comporte des informations divergentes de celles des référentiels, une coloration rouge apparaît et la sauvegarde de la modification est empêchée.

On peut également effectuer :

- Des corrections en série grâce à la catégorie *similar*
- Des modifications sur le texte d'entrée avec la catégorie "+", si ce dernier s'avère fautif :
  - modification de la forme fautive ;
  - suppression du token ;
  - ajout d'un token.

# 2.2 Fonctionnalités avancées : corrections en fonction de filtres de recherche




*Interface de requêtes, capture d'écran, copyright © 2018 Clérice, Pilla, & Camps*

## Vérifier son annotation

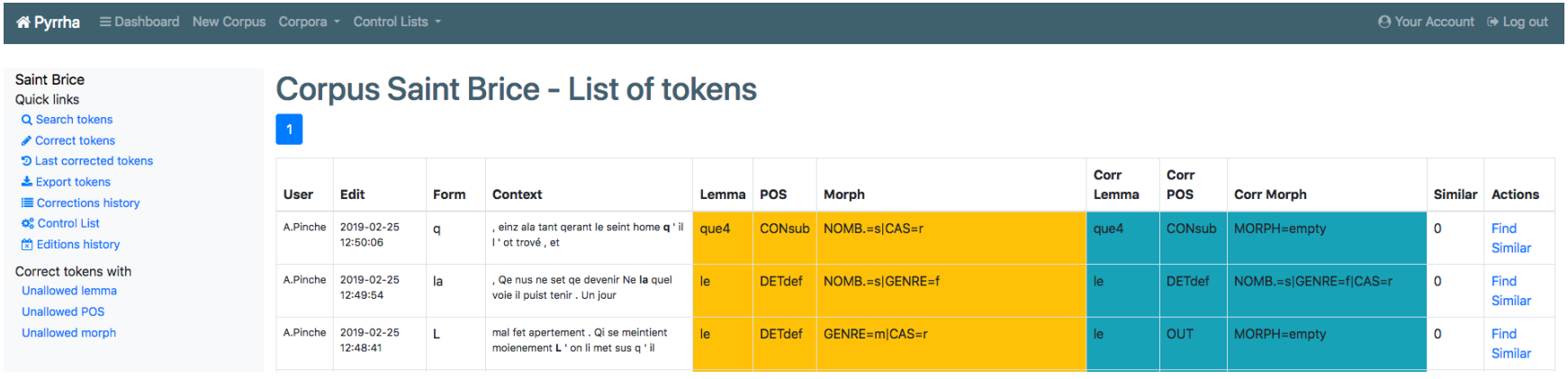
- Chercher un token en fonction de :
  - sa forme ;
  - son lemme ;
  - sa POS ;
  - son étiquetage morphosyntaxique.

## Nettoyer son annotation

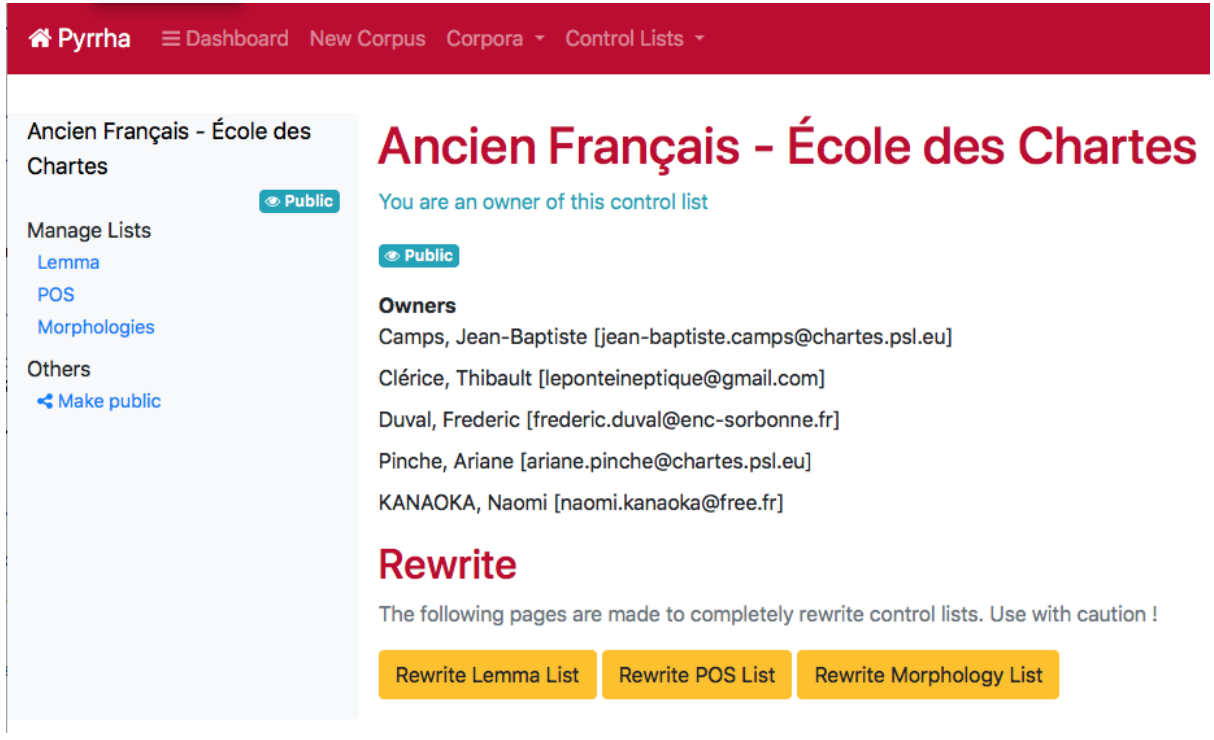
- Accéder aux formes qui possèdent un étiquetage qui ne correspond pas au référentiel choisi :
  - Unallowed lemma;
  - Unallowed POS ;
  - Unallowed morph.

# Contrôle des corrections opérées

## Historique des corrections



# Accès aux listes de contrôle




Les listes de contrôle sont ordonnées selon trois catégories : Lemma, POS, morphologie. Toutes modifications sur les listes officielles sont soumises à modération.

## 2.3 Exports des données

A l'issu des corrections, les données peuvent être intégralement sauvegardées et exportées dans un fichier CSV au format Pie ou en XML TEI pour être interrogées en vue d'analyses statistiques.

## Exemple d'encodage XML TEI

```
<w xml:id=" t1" n=" 1" lemma=" ci" type=" POS=ADVgen|DEGRE=-" >Ci</w>
```

1. @xml:id pour l'identifiant du mot qui correspond dans la plupart de cas à l'identifiant du mot dans l'interface Pyrrha, soit son numéro, précédé de t.
2. @n correspond à l'id du token dans l'interface
3. @lemma qui correspond au lemme
4. @POS qui correspond à la fois à la catégorie POS et à la catégorie Morph.

## 3. Exploitation des données : deux cas pratiques

L'annotation a été générée à partir d'une lemmatisation automatique par Pie entraîné sur le modèle pour l'ancien français et corrigé au moyen de l'interface Pyrrha.

## 3.1 Réalisation de yod + ATA > pic. -ie (franc. -iée)

- Le texte possède 169 participes passés au féminin employés avec l'auxiliaire être
- Les 136 formes ne se terminant pas en -ie présentent la marque flexionnelle du féminin :

  abatue, acostumee, adolee, ajostee, alee, aombree, aoree, aornee, apelee, arivees, assamblez, assise, atornee, avenue, avironnee, beneïe, brullee, celee, chantee, cheüe, consomee, contee, corrompue, coverte, creüe, curee, dampnee, delivree, demenee, dervee, desconfortee, desfendue, desmesuree, destruite, detrete, devenue, doblee, dontee, dounee, enclose, encortinee, enluminee, entree, escrite, espandue, esparse, esprise, esteinte, etc.

- 32 formes se terminent en "-ie".

  *acouchie, aemplie, aidie, amenuisie, apareillie, apeisie, apesie, aprochie, baillie, convertie, deguerpie, departie, enforcie, esmaie, essaucie, establie, florie, garantie, garie, guerie, mollie, negie, noncie, raemplie, raverdie, trenchie*

- 1 seule forme se termine en "-iee" : sechiee

## 3.2 Alternance graphique à l'initiale de c et ch

### 3.2.1 c+e/i

- 1432 occurrences de termes avec un lemme commençant par "ce"
  - 1195 occurrences, soit 85% des cas, sont ces occurrences de démonstratifs (ce, cest, cel)
  - 3 termes présentent une initiale en ch : chainte (1), cheinture (2)
- 98 occurrences de termes avec un lemme commençant par "che", aucune variation observée entre la graphie du lemme et du terme
- 33 termes commencent par chi, mais une seule occurrence présente une graphie en ci, *cier* pour chier

### 3.2.2. c+a

- 600 occurrences de termes avec un lemme commençant par "ca"
  - 4 occurrences avec une forme ne ch initial : chartage (1) pour Cartage, Chaton (2) pour Caton, chariole (1) pour cariole
- 183 occurrences de termes avec un lemme commençant par "cha"
  - 1 occurrence de cascun (sur 28 occurences du même terme)
  - 7 occurrences de ceaille pour chaille (sur 8 occurences du même terme)

# Bibliographie sélective


- Clérice, T., Pilla, J., & Camps, J.-B. (2018). hipster-philology/pyrrha: 1.0.1. https://doi.org/10.5281/zenodo.2325428
- Dees, A., Dekker, M., Huber, O., & Van Reenen-Stein, K. (1987). Atlas des formes linguistiques des textes littéraires de l'ancien français (Reprint 2014).
- Gossen, C. T. (1951). Petite grammaire de l'ancien picard : phonétique, morphologie, syntaxe, anthologie et glossaire. Paris: C. Klincksieck.

- Guillot, C., Prévost, S., & Lavrentiev, A. (2013, avril 8). Manuel de référence du jeu Cattex09.
- Manjavacas, E., Kestemont, M., & Clérice, T. (2019). emanjavacas/pie v0.1.0. https://doi.org/10.5281/zenodo.1637878
- Mellet, S. (2002). La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ? Médiévales, 21 (42), 13-26.