



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

Ariane Pinche

## ► To cite this version:

Ariane Pinche. Annoter facilement un corpus complexe: l'exemple de Pyrrha, interface de post-correction, et Pie, lemmatiseur et tagueur, pour l'ancien français. Rencontres lyonnaises des jeunes chercheurs en linguistique historique, Jun 2019, Lyon, France. hal-02182740

HAL Id: hal-02182740

<https://hal.archives-ouvertes.fr/hal-02182740>

Submitted on 16 Jul 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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](http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf)

## Exemples d'annotation

Form	Lemma	POS	Morph
commence	comencier	VERcjk	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 ;

## Create a new corpus

**Metadata**

Corpus Name

This should be a clear name

Left Context

Number of words to display on the left of the word to annotate

Right Context

Number of words to display on the right of the word to annotate

Delimiter token

[Optional] A specific token that is used as a delimiter between passages.

**Data**

Tokens (as TSV content)

The TSV should at least have the headers : form, lemma, POS, morph

Example :

form	lemma	POS	morph
SOIGNORS	seignor	NOMcom	NOMB.=p GENRE=m CAS=n
or	or4	ADVgen	DEGRE=-
escoutez	escouter	VERcJg	MODE=imp PERS.=2 NOMB.=p

Lemmatize

If your text is not lemmatized, select the language and click on lemmatize.

Tokenize (beta)

If your text is not tokenized and you don't need to pre-lemmatize it, you can use this function

Remove hyphens (Be careful with this function)

Keep punctuation

**Control Lists**

Use an existing control list

By using shared control lists, you ensure that you stick to accepted values in the academic community. You will be able to propose new values to the administrators of control lists.

Write your own

In case the configurations provided do not fit your need, you can create your own configuration. You will be able to share it with collaborators, propose it as a canon setting for the whole base of users. You can also later add bibliographic information about your settings.



- 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

**Data**

Tokens (as TSV content)

The TSV should at least have the headers : lemma, POS, morph, form

Lemmatize

If your text is not lemmatized, select the language and click on lemmatize.

Ancien Français Lemmatize

Tokenize (beta)

If your text is not tokenized and you don't need to pre-lemmatize it, you can use this function

Remove hyphens (Be careful with this function)

Keep punctuation

Tokenize

*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

Pyrrha Dashboard New Corpus Corpora Control Lists Your Account Log out

Saint Brice  
Quick links  
Search tokens  
Correct tokens  
Last corrected tokens  
Export tokens  
Corrections history  
Control List  
Editions history  
Correct tokens with  
Unallowed lemma  
Unallowed POS  
Unallowed morph

### Corpus Saint Brice - List of tokens

1 2 3 4 5 ... 17 18

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

Pyrrha Dashboard New Corpus Corpora Control Lists Your Account Log out

Saint Brice  
Quick links  
Search tokens  
Correct tokens  
Last corrected tokens  
Export tokens  
Corrections history  
Control List  
Editions history

Correct tokens with  
Unallowed lemma  
Unallowed POS  
Unallowed morph

### Corpus Saint Brice - Search tokens within the corpus

Form	Lemma	POS	Morph
<input type="text" value="Brice"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

\* can be used to match partial words, eg. **ADV\***  
! can be used to negate a match, eg. **!PRE**  
| can be used to perform an OR operation, eg. **s\*t | s\*s**  
To search forms which do not contain 'e': **! \*e\***

[Search](#)

8 matches found

1

Id	Form	Lemma	POS	Morph	Context	Similar	Save	+
9	Brice	Brice	NOMpro	NOMB.=s GENRE=m CAS=r	Ci commence la vie de mon seigneur seint <b>Brice</b> . Quant seinz Brices estoit jovenceaus , il		Save	+
350	Brice	Brice	NOMpro	NOMB.=s GENRE=m CAS=r	enfermeté . Et seinz Martins vint a seint <b>Brice</b> qj diacres estoit et parla en tel maniere		Save	+

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

Pyrrha Dashboard New Corpus Corpora Control Lists Your Account Log out

Saint Brice  
Quick links  
Search tokens  
Correct tokens  
Last corrected tokens  
Export tokens  
Corrections history  
Control List  
Editions history

Correct tokens with  
Unallowed lemma  
Unallowed POS  
Unallowed morph

### Corpus Saint Brice - List of tokens

1

User	Edit	Form	Context	Lemma	POS	Morph	Corr Lemma	Corr POS	Corr Morph	Similar	Actions
A.Pinche	2019-02-25 12:50:06	q	, einz ala tant qerant le seint home q ' il l ' ot trové , et	que4	CONsub	NOMB.=s CAS=r	que4	CONsub	MORPH=empty	0	Find Similar
A.Pinche	2019-02-25 12:49:54	la	, Qe nus ne set qe devenir Ne la quel voie il puist tenir . Un jour	le	DETdef	NOMB.=s GENRE=f	le	DETdef	NOMB.=s GENRE=f CAS=r	0	Find Similar
A.Pinche	2019-02-25 12:48:41	L	mal fet apertement . Qi se meintient moienement L ' on li met sus q ' il	le	DETdef	GENRE=m CAS=r	le	OUT	MORPH=empty	0	Find Similar

# Accès aux listes de contrôle

Pyrrha Dashboard New Corpus Corpora Control Lists

Ancien Français - École des Chartes

Public

You are an owner of this control list

Public

**Owners**

- Camps, Jean-Baptiste [jean-baptiste.camps@chartes.psl.eu]
- Clérice, Thibault [leponteineptique@gmail.com]
- Duval, Frederic [frederic.duval@enc-sorbonne.fr]
- Pinche, Ariane [ariane.pinche@chartes.psl.eu]
- KANAOKA, Naomi [naomi.kanaoka@free.fr]

**Rewrite**

The following pages are made to completely rewrite control lists. Use with caution !

Rewrite Lemma List Rewrite POS List Rewrite Morphology List

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'issue 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, assemblez, assise, atornee, avenue, avironnee, beneie, 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, escrete, expandue, esparsse, esprise, esteinte, etc.



- 32 formes se terminent en "-ie".

*acouchie, aemplie, aidie, amenuisie, apareillie, apeisie, apesie, aprochie, baillie, convertie, deguerpie, departie, enforcie, esmaie, essaucie, estableie, 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 occurrences du même terme)
  - 7 occurrences de ceaille pour chaille (sur 8 occurrences 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.