

Phyrra : lemmatisation et correction

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

Vincent Jolivet, École nationale des chartes, Centre Jean Mabillon

Thibault Clérice, École nationale des chartes et Univ. Jean Moulin Lyon 3,
Centre Jean Mabillon

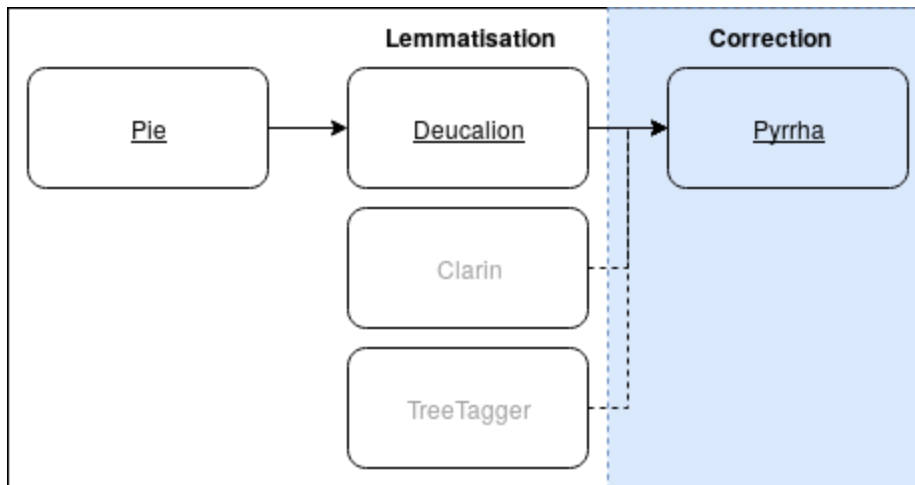
Journée d'initiation à la lemmatisation des textes médiévaux, 6 juin 2019

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

I - Lemmatisation d'un corpus

1.1 - Infrastructure



Pie, Deucalion et Pyrrha : Open-Source et Python

1.2 - Pie et Deucalion

Pie est un outil qui s'entraîne sur des corpus, et s'évalue sur un corpus non connu et déjà annoté. Il est écrit en Python et peut s'utiliser en ligne de commande.

Deucalion est son interface web.

Manjavacas, E., Kádár, A., & Kestemont, M. (2019). *Improving Lemmatization of Non-Standard Languages with Joint Learning*, NAACL, <https://arxiv.org/abs/1903.06939>

Comparaison :

	Full	Ambiguous	Unknown
Morphette/Lemming	91.11	91.79	35.48
Pie	94.0	92.81	65.39

1.3 - Les modèles d'annotation de Pyrrha

Pyrrha est un outil de correction en ligne. Il peut être utilisé de concert avec Deucalion pour obtenir des textes lemmatisés, mais peut tout aussi bien recevoir des données d'autres outils, du moment qu'elles respectent un format TSV bien précis (Excel/Libre Office Calc -> Export CSV, delimitateur : tabulation).

- 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://lindat.mff.cuni.cz/services/udpipe/>

1.4 - Exemples d'annotation

Form	Lemma	POS	Morph
commence	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

1.5 - Résultats sur Deucalion Latin LASLA

Testé sur 141.137 mots sur un nombre de phrases non-compté.

	Lem	POS	Voix	Mode	Deg	Nomb	Pers	Tps	Cas	Gen
Tous	97.29	96.55	99.18	98.36	98.3	97.88	99.18	98.75	93.74	97.27
Ambigus	92.51	91.70	94.53	88.95	92.66	92.25	92.47	91.91	85.96	89.58
Formes Inconnues	85.74	89.66	95.62	93.38	94.52	95.67	97.50	94.93	89.13	92.83
Lemmes Inconnus	61.59	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

1.6 - Résultats sur Deucalion Ancien français

Testé sur 17.751 mots pour la morphologie, 48.317 pour lemme et POS.

Morphologie: 1286 mots inconnus en tests, 2367 ambigus

	Lemmes	POS	Mode	Degré	Nombre	Personne	Temps	Cas	Genre
Tous	96.38	96.13	98.74	98.26	95.86	97.86	98.55	93.38	95.18
Ambigus	96.65	95.49	95.9	95.02	94.29	95.99	95.77	91.64	93.24
Formes Inconnues	64.29	86.77	94.01	94.32	89.97	93.62	92.77	85.15	84.91
Lemmes Inconnus	72.90	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

1.7 - Deucalion via Docker

Il est possible d'installer simplement une version de Deucalion sur son propre PC via Deucalion via Docker ou Kitematic.

Latin

<https://hub.docker.com/r/ponteineptique/deucalion-model-lasla>

Ancien Francais

<https://hub.docker.com/r/ponteineptique/deucalion-model-af>

II - 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.

Tutoriel : <https://github.com/hipster-philology/pyrrha/blob/dev/docs/fr/tutoriel.md>

Scénario d'usage

1. [Créer un nouveau corpus](#) : un utilisateur crée un corpus à corriger.
2. [Gérer le travail collaboratif](#) : cet utilisateur peut inviter d'autres utilisateurs à collaborer à la correction du corpus.
3. [Corriger l'annotation](#).
4. [Exporter les données corrigées](#).

II.1 - Gérer son compte utilisateur

1.1. Créer son compte

Lien **Register** : <https://dev.chartes.psl.eu/pyrrha/account/register>

1. Renseigner le formulaire.
2. Confirmer l'inscription en cliquant sur le lien reçu dans sa messagerie.

1.2. Modifier son compte

Onglet **Your Account**

Possibilité de mettre à jour l'adresse mail et le mot de passe.

1.3. Supprimer son compte

TODO

II.2. Gérer ses corpus

2.1. Créer un corpus

Onglet [New Corpus](#)

Un *nouveau corpus* est un texte étiqueté que l'on souhaite corriger.

À sa création, il convient donc de lui [associer les ressources \(listes de contrôle\)](#) utiles à la reprise de l'étiquetage.

2.1.1. Importer un texte

- **Metadata > Corpus Name** : nommer explicitement le nouveau corpus pour faciliter le suivi de nombreux projets.
- **< Metadata > Left and right context** : définir la taille des contextes gauche et droit autour du token éditable dans l'interface de correction (3 mots par défaut de part et d'autre du token).
- **Data > Tokens (as TSV content)** : copier-coller le texte étiqueté à corriger au format **TSV**, en respectant l'en-tête suivante :

```
Form      Lemma    POS      Morph
son       son4      DETpos   PERS.=3|NOMB.=s|GENRE=m|CAS=r
seigneur  seignor   NOMcom   NOMB.=s|GENRE=m|CAS=r
voit      v  oir    VERCjg   MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s
bien      bien1     ADVgen   DEGRE=p
...
```

- Penser à cliquer en bas de page sur le bouton **Submit** pour enregistrer le nouveau corpus.

2.1.2. Tokenizer un nouveau corpus (beta)

Si le texte n'est pas encore étiqueté, il est possible d'importer simplement le texte brut :

- Copier-coller votre texte dans le champs `Data > Tokens (as TSV content)` .
- Cliquer sur le bouton `Tokenize` .

Le texte est reformaté pour les besoins de l'annotation : chaque token est inscrit en début de ligne et l'en-tête obligatoire (`form | lemma | POS morph`) est ajoutée.

Data

Tokens (as TSV content)

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

Lemmatize

Ancien Français

Lemmatize

Tokenize (beta)

☐ Remove hyphens (Be careful with this function)

☒ Keep punctuation

Tokenize

À la création du nouveau corpus, l'École des chartes propose un service de lemmatisation pour l'ancien français et le latin.

- Copier-coller le texte dans le champs `Data > Tokens (as TSV content)` .
- Dans le menu déroulant, sélectionner le modèle de langue.
- Cliquer sur le bouton `Lemmatize` .

The screenshot shows the Pie web interface for lemmatization. At the top, there's a 'Data' section with a text area labeled 'Tokens (as TSV content)' and a note: 'The TSV should at least have the headers : lemma, POS, morph, form'. Below this is a 'Lemmatize' section with a dropdown menu currently set to 'Ancien Français' and a 'Lemmatize' button. A note below the dropdown says: 'If your text is not lemmatized, select the language and click on lemmatize.' To the right of the dropdown is a 'Tokenize' button. Below the 'Tokenize' button is a 'Tokenize (beta)' section with a note: 'If your text is not tokenized and you don't need to pre-lemmatize it, you can use this function'. There are two checkboxes: 'Remove hyphens (Be careful with this function)' which is unchecked, and 'Keep punctuation' which is checked.

Le service de lemmatisation utilise des modèles [Pie](#) (Manjavacas, E., Kestemont, M., & Clérice, T. (2019). `emanjavacas/pie v0.1.0`.

<https://doi.org/10.5281/zenodo.1637878>) :

- Deucalion pour le latin (<https://doi.org/10.5281/zenodo.2707476>) :
 - modèle entraîné sur les données du [LASLA](#).
- Deucalion pour l'ancien français (<https://doi.org/10.5281/zenodo.3237455>) :
 - lemmes issus du Tobler-Lommatzsch ;

2.1.4. Associer des listes de contrôle

`Control Lists` . Les listes de contrôle facilitent la correction de l'étiquetage : elles permettent d'isoler les étiquettes non autorisées ou inconnues et d'encadrer la saisie du correcteur (suggestions et autocomplétion).

- Liste de lemmes (`Lemma List`)
- Liste des étiquettes grammaticales (`POS List`)
- Liste des étiquettes morphologiques (`Morph List`)

Cocher (au choix) :

- `Use an existing control list` pour utiliser des listes prédéfinies et partagées (et y contribuer). Ces listes sont disponibles pour :
 - l'ancien Français ;
 - le français moderne ;
 - le latin ([LASLA](#)).
- `Write your own` pour créer ses propres listes si aucune des listes partagées ne convient au besoin.

2.3. Collaborer

Onglet `Dashboard > Corpora > corpus_name`

Il est possible d'inviter des utilisateurs enregistrés à collaborer à la correction d'un corpus.

Inviter des utilisateurs

- Dans la liste `Grant access to a user`, cliquer sur les utilisateurs invités : ils s'ajoutent à la liste des utilisateurs associés au corpus (liste `View and manage corpus users`).
- Pour associer certains utilisateurs en tant qu'administrateur du corpus, cocher la case `Owner`.
- Cliquer en bas de page sur le bouton `Save modifications`.

Retirer un utilisateur de la liste associée au corpus

- Dans la liste `View and manage corpus users`, cliquer sur l'icône `Corbeille`.
- Cliquer en bas de page sur le bouton `Save modifications`.

III.3 - Corriger les données

3.1. Fonctionnalités de base : relecture et édition des corrections

Onglet `Corpora > corpus_name > Edit tokens`

ou `Quick links > Correct tokens`

L'interface affiche un tableau à 9 colonnes, dont 3 sont éditables :

1. `Id` : identifiant attribué à chaque token (mots et éléments de ponctuation) ;
2. `Form` : **éditable**, terme tel qu'il apparaît dans le texte ;
3. `Lemma` : **éditable**, lemme attribué à chaque token ;
4. `POS` : **éditable**, étiquette grammaticale du token ;
5. `Morph` : étiquette morpho-syntaxique du token ;
6. `Context` : le token en contexte ([configurer le contexte](#)) ;
7. `Similar` : nombre de token similaires (pour les [corrections par lots](#)) ;
8. `Save` : sauvegarder les modifications ;
9. `+` : options de modification du token : correction, suppression, ajout.

3.2. Corriger les étiquettes **Lemma** , **Pos** et **Morph**

Quick links > Correct tokens

1. Cliquer dans la cellule à corriger.
2. Corriger la valeur.
3. Cliquer sur **save** pour enregistrer la modification.

Gestion des erreurs. Si la valeur saisie est absente de la liste de contrôle correspondante, la cellule apparaît en rouge et la sauvegarde est empêchée.

Id	Form	Lemma	POS	Morph	Context	Similar	Save	+
1	D'autres	dature	ADJqua	NOMB.=p GENRE=m CAS=n	D'autres gens sonjier me veil	2	Save	+
2	gens	gent1	NOMcom	NOMB.=p GENRE=m CAS=n	D'autres gens sonjier me veil qui	4	Save	+
3	sonjier	songier	VERinf	NOMB.=x GENRE=x	D'autres gens sonjier me veil qui sont	0	Save	+
4	me	je	PROper	PERS.=1 NOMB.=s GENRE=m CAS=i	D'autres gens sonjier me veil qui sont fieres	0	Save	+

Si nécessaire (définition d'un nouveau lemme par ex.), il est possible de [modifier les listes de contrôle](#).

- Match :
 - Partial :
 - Complete :
- Match at least
 - Lemma :
 - POS :
 - Morph :
- Different on
 - Lemma :
 - POS :
 - Morph :

Corpus Monstres - List of tokens

PSL							
							
Id	Form	Lemma	POS	Morph	Context	Similar	Save +
1	D'autres	dature	ADJqua	NOMB,=p GENRE=m CAS=n	D'autres gens sonjier me veil	2	Save +
2	gens	gent1	NOMcom	NOMB,=p GENRE=m CAS=n	D'autres gens sonjier me veil qui	3	Save +

3.4. Corriger par lots grâce aux filtres de recherche

Quick links > Search tokens

1. Rechercher des tokens selon :

- leur forme (**Form**) ;
- et/ou leur lemme (**Lemma**) ;
- et/ou leur POS (**POS**) ;
- et/ou leur étiquette morpho-syntaxique (**Morph**).
- NB. les expressions régulières (Regex) sont autorisées.

1. Corriger par lots.

3.5. Contrôler et nettoyer l'annotation

Menu `Correct tokens with`

Ce raccourci permet de lister les tokens dont l'étiquetage n'est pas validé par les listes de contrôle.

1. Cliquer sur :

- `Unallowed lemma` : liste des tokens dont le lemme est inconnu.
- `Unallowed POS` : liste des tokens dont l'étiquette grammaticale est inconnue.
- `Unallowed morph` : liste des tokens dont l'étiquette morpho-syntaxique est inconnue.

1. Corriger par lot ou modifier la liste de contrôle.

3.6. Retourner à sa dernière correction

Quick links > Last corrected tokens

3.7. Suivre les corrections de l'annotation

Quick links > Corrections history

3.8. Corriger le texte annoté

Si nécessaire, vous pouvez corriger le texte annoté (la liste des tokens) grâce au raccourci `+` de la dernière colonne qui ouvre un menu contextuel :

- `Edit the form` : modification de la forme fautive ;
- `Delete the row` : suppression du token ;
- `Add a token after this one` : ajout d'un token.

Pour suivre les modifications apportées au texte annoté, cliquer sur `Quick links >`
`Editions history`.

III.4 - Modifier les listes de contrôle

Quick links > Control Lists

Utilisation des listes partagées (public list)

Les modifications sur ces listes sont soumises à modération.

Utilisation des listes personnalisées (private list)

TODO

III.5 - Exporter les données

Quick links > Export tokens

À tout moment, les données peuvent être intégralement exportées en TSV ou en XML/TEI.

Pandora/Pie CSV

Export TSV avec l'en-tête Form | Lemma | POS | Morph .

TEI

```
<w xml:id="t6" n="6" lemma="qui" type="POS=PROadv|NOMB.=s|GENRE=m|CAS=n">qui</w>
```

- @xml:id : identifiant Pyrrha du token (numéro d'ordre dans le corpus)
- @lemma : lemme
- @type : concaténation des étiquettes POS et morpho-syntaxique.

Contacts et adresses

- Bug ? Fonctionnalité ?
 - Si possible : <https://github.com/hipster-philology/pyrrha/issues>
 - Sinon :
 - julien.pilla@chartes.psl.eu
 - vincent.jolivet@chartes.psl.eu
 - thibault.clerice@chartes.psl.eu
- Fournir des données d'entraînement:
 - thibault.clerice@chartes.psl.eu
 - jean-baptiste.camps@chartes.psl.eu

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.