Lemmatiser des textes et corriger l'annotation grâce à l'apprentissage profond et Pyrrha

Pyrrha (suite)

Thibault Clérice[1,2] Matthias Gille Levenson[3] Lucence Ing[1]
Ariane Pinche[1,2] Simon Gabay[4] Jean-Baptiste Camps[1]

1. École nationale des chartes, PSL
2. Université Lyon
3. Casa de Velázquez
4. Université de Genève

Humanistica 2021
9-12 mai 2021

1 Référentiels

2 Custom Dictionary

3 Collaboration et exports

Utilité des référentiels

Intérêts généraux

- création de données renvoyant à des dictionnaires et étiquettes de référence
- validation ou non de données au sein du travail

Utilité des référentiels

Intérêts généraux

- création de données renvoyant à des dictionnaires et étiquettes de référence
- validation ou non de données au sein du travail

Avantages au sein des différentes étapes de travail

- uniformisation de ses données
- travail collaboratif facilité
- partage et réutilisation des données

Les listes de contrôle dans Pyrrha (1)

Quand on crée un corpus

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.

Ancien Français - École des Chartes

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.

Submit

- quatre listes disponibles : latin classique, ancien français, ancien occitan, français moderne
- possibilité de créer ses propres listes

Les listes de contrôle dans Pyrrha (2)

Les listes peuvent définir :

- les lemmes
- les parties du discours
- la morphologie

Les listes s'enrichissent au fur et à mesure des utilisations :

- noms propres
- nouveaux mots rencontrés dans un texte, absents des référentiels originaux

Référentiel : latin classique

Classical Latin (LASLA derived)

- lemmes : *Lexicon totius latinitatis* de Forcellini, éd. de Corradini, Padoue, 1864
- parties du discours : Thibault Clérice d'après les travaux du LASLA de D. Longrée et al.
- morphologie : Thibault Clérice d'après les travaux du LASLA de D. Longrée et al.

Documentation : `https://github.com/lascivaroma/forcellini-lemmas/`

Référentiel : ancien français

Ancien français (École des chartes)

- lemmes : Tobler-Lommatzsch, *Altfranzösisches Wörterbuch*
- parties du discours : CATTEX2009, `http://bfm.ens-lyon.fr/spip.php?article176`, de la *Base de français médiéval*
- morphologie : CATTEX2009

Documentation : `https://github.com/Jean-Baptiste-Camps/Geste/wiki`

Référentiel : ancien occitan

Ancien occitan

- lemmes : Dictionnaire de l'occitan médiéval (DOM), `http://www.dom-en-ligne.de/index.html`
- parties du discours : CATTEX2009
- morphologie : CATTEX2009

Référentiel : français moderne

Français - LGeRM (modernisé et non modernisé)

- lemmes : LGeRM MODE, `http://stella.atilf.fr/LGeRM/` de l'Atilf
- parties du discours : CATTEX2009
- morphologie : CATTEX2009

Documentation : `https://github.com/e-ditiones/LEM17`

Documentation des listes

Control Lists/*liste*/Guidelines

Pyrrha | Dashboard | New Corpus | Control Lists | Bug or request ? | Help | Your Acc

Classical Latin (LASLA-Derived)

Public

Guidelines

Manage Lists

Lemma

POS

Morphologies

Others

Propose changes

Guidelines

Lemmas

You can find the lemma dictionary there : https://lascivaroma.github.io/forcellini-lemmas/index.html. You can download the TSV here https://github.com/lasci lemmas/raw/master/dictionnary.tsv

Foreign Words

- Greek words need to be lemmatized as such: `Lemma: [Greek]`, `POS: FOR` (Foreign), `Morph: Morph=Empty`
- Old French words need to be lemmatized as such: `Lemma: [FRO]`, `POS: FOR` (Foreign), `Morph: Morph=Empty`

POS spécifique pour lemmes spécifiques

Les parties suivantes listes les POS les plus complexes, avec des exemples.

ADJadv.ord

primo, primum, prius, quarto, quartum, quintum, secundo, sextum, tertio, tertium

ADJdis

15, 21, 60, binus, centenus, denus, ducenus, duodeni, duodenus, nonagenus, nouenus, octonus, quadragenus, quadrinus, quartus, quaternus, quingenus, quinqu senus, septenus, sescenus, sexagenus, sexcenus, singulus, terni, ternus, trecentenus, trecenus, tricenus, trinus, uicenus, undenus

Suggestions de modifications

`Dashboard/`*corpusVoulu*`/Control List/Propose changes`

Ancien Français - École des Chartes

Public

Guidelines

Manage Lists

Lemma

POS

Morphologies

Others

Propose changes

Contact control lists administrators

This page is built so that you can contact administrators of a control list to propose ne
email so that discussion could be rich when needed.

Title | Title

Message

Message

Send mail

Écrire sa propre liste

Quand on crée un corpus

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.

Load a configuration

This helps kickstarting configuration

Ancien Français | Load

Allowed lemma

This should be formatted as a list of lemma separated by new line

Allowed POS

This should be formatted as a list of POS separated by comma **and no space**

Allowed Morph (as TSV content)

The TSV should at least have the header : label and could have a `readable` column for human

Submit

Gérer sa liste

`Control Lists/`*`son référentiel`*

Control List test

Private

Guidelines

Manage Lists

Lemma

POS

Morphologies

Others

Propose changes

Rename

Edit informations

Control List test

Description

Bibliography

Information

You are an owner of this control list

Private

Owners

Ing, Lucence [lucence.ing@chartes.psl.eu]

Rewrite

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

Rewrite Lemma List | Rewrite POS List | Rewrite Morphology List

1 Référentiels

2 Custom Dictionary

3 Collaboration et exports

Une nouvelle fonctionnalité

→ Nouvelle fonctionnalité pour pallier à la longueur de la démarche (nécessaire !) du processus de demande d'ajout d'une étiquette.

Les dictionnaires personnalisés :

- dépendent d'un corpus particulier
- permettent une fluidité dans le travail
- ne cassent pas la conformité avec les référentiels

Custom Dictionary

Sous les Preferences, sous le nom du corpus

Add10293

- Preferences
- Custom Dictionary

Quick links

- Statistics
- Search tokens
- Correct tokens
- Bookmark
- Export tokens
- Corrections history
- Control List
- Editions history
- Switch control list
- Delete the corpus

Correct tokens with

- Unallowed lemma
- Unallowed POS

Custom dictionary

Dictionaries

Lemma

Lancelot

List of supplementaries lemma for this corpus. One per line

Morph

List of supplementaries morph for this corpus. One per line, a human readable version of the c
transformed into tabulation). eg.

```
Morph=MyMorph    My human readable morph
```

POS

Ajouter une entrée dans le dictionnaire

Quand on entre une valeur non reconnue

					pensé ; si			
2524	Claudas	Claudas	NOMpro	NOMB.=s|GENRE=m|CAS=r	. Quant vint a l ' endemain que li senescaus **Claudas** ala querre les enfans , encore n ' avoit Lyoniax		Save	
					Invalid value in lemma	Add `Claudas` to custom dictionary of *lemma*		

1 Référentiels

2 Custom Dictionary

3 Collaboration et exports

Partager un corpus

Dashboard/*Corpus*

Back to dashboard

View and manage corpus users

Add or remove corpus access to users

Add10293

ID	First name	Last name	Owner	
2	Lucence	Ing	✓	🗑

Grant access to a user

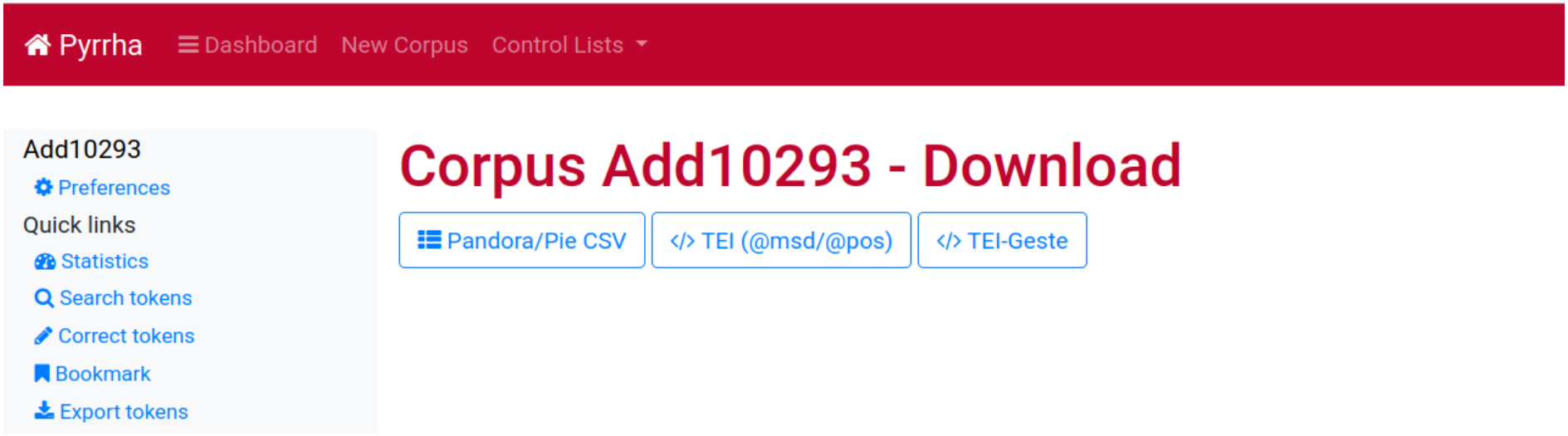
The user will be able to use the post-correction tools but also grant and remove accesses to this corpus

All account types | Search users...

- rechercher des personnes usagères et leur partager le corpus
- supprimer l'accès à des personnes

Exporter les données

Dans le menu : `Export tokens`



3 types d'export disponibles : un en format CSV, deux en format TEI

Exporter en CSV

A	B	C	D
form	lemma	POS	morph
Tant	tant	ADVgen	DEGRE=-
m	je	PROper	PERS.=1|NOMB.=s|GENRE=f|CAS=r
'	'	PONfbl	MORPH=empty
avés	avoir	VERcjg	MODE=ind|TEMPS=pst|PERS.=2|NOMB.=p
conjuree	conjurer	VERppe	NOMB.=s|GENRE=f|CAS=r
,	,	PONfbl	MORPH=empty
fait	faire	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s
la	le	DETdef	NOMB.=s|GENRE=f|CAS=n
damoisele	damoisele	NOMcom	NOMB.=s|GENRE=f|CAS=n
,	,	PONfbl	MORPH=empty
que	que4	CONsub	MORPH=empty
ja	ja	ADVgen	DEGRE=-
plus	plus	ADVgen	DEGRE=-
ne	ne1	ADVneg	MORPH=empty
vous	vos	PROper	PERS.=2|NOMB.=p|GENRE=m|CAS=i
iert	estre1	VERcjg	MODE=ind|TEMPS=fut|PERS.=3|NOMB.=s
chelei	celer1	VERcjg	NOMB.=s|GENRE=m|CAS=n

Exporter en TEI


```
−<text xml:lang="fr">
 −<body xml:lang="fro">
  −<div>
   −<ab>
     <w xml:id="t1" n="1" lemma="tant" pos="ADVgen" msd="DEGRE=-">Tant</w>
     <w xml:id="t2" n="2" lemma="je" pos="PROper" msd="PERS.=1|NOMB.=s|GENRE=f|CAS=r">m</w>
     <w xml:id="t3" n="3" lemma="'" pos="PONfbl" msd="MORPH=empty">'</w>
     <w xml:id="t4" n="4" lemma="avoir" pos="VERcjg" msd="MODE=ind|TEMPS=pst|PERS.=2|NOMB.=p">avés</w>
     <w xml:id="t5" n="5" lemma="conjurer" pos="VERppe" msd="NOMB.=s|GENRE=f|CAS=r">conjuree</w>
     <w xml:id="t6" n="6" lemma="," pos="PONfbl" msd="MORPH=empty">,</w>
     <w xml:id="t7" n="7" lemma="faire" pos="VERcjg" msd="MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s">fait</w>
     <w xml:id="t8" n="8" lemma="le" pos="DETdef" msd="NOMB.=s|GENRE=f|CAS=n">la</w>
     <w xml:id="t9" n="9" lemma="damoisele" pos="NOMcom" msd="NOMB.=s|GENRE=f|CAS=n">damoisele</w>
     <w xml:id="t10" n="10" lemma="," pos="PONfbl" msd="MORPH=empty">,</w>
     <w xml:id="t11" n="11" lemma="que4" pos="CONsub" msd="MORPH=empty">que</w>
     <w xml:id="t12" n="12" lemma="ja" pos="ADVgen" msd="DEGRE=-">ja</w>
     <w xml:id="t13" n="13" lemma="plus" pos="ADVgen" msd="DEGRE=-">plus</w>
     <w xml:id="t14" n="14" lemma="ne1" pos="ADVneg" msd="MORPH=empty">ne</w>
     <w xml:id="t15" n="15" lemma="vos" pos="PROper" msd="PERS.=2|NOMB.=p|GENRE=m|CAS=i">vous</w>
     <w xml:id="t16" n="16" lemma="estre1" pos="VERcjg" msd="MODE=ind|TEMPS=fut|PERS.=3|NOMB.=s">iert</w>
     <w xml:id="t17" n="17" lemma="celer1" pos="VERcjg" msd="NOMB.=s|GENRE=m|CAS=n">chelei</w>
```

Zoom sur les données d'une occurrence en TEI

```
<w xml:id="t4" n="4" lemma="avoir" pos="VERcjg" msd="MODE=ind|
TEMPS=pst|PERS.=2|NOMB.=p">avés</w>
```

`<w>` : la balise contenant l'occurrence et ses métadonnées
`@xml:id` : un attribut contenant l'identifiant de l'occurrence
`@n` : un attribut contenant le numéro de l'occurrence
`@lemma` : un attribut contenant le lemme de l'occurrence
`@pos` : un attribut contenant la partie du discours de l'occurrence
`@msd` : un attribut contenant l'étiquette morphologique de l'occurrence