

Deucalion et Pyrrha

Environnement pour la lemmatisation et la post-correction à l'École des chartes

Text Encoding: Latinists looking for new synergies, LASLA, Liège,
8-9 novembre 2018

Le Projet

Source

- Thèse d'A. Pinche (Lyon 3 et ENC) et travail de J.-B. Camps (ENC) sur des textes en Ancien Français.
- Besoin d'accélération du travail de lemmatisation pour des besoins de recherches
 - Création d'index pour une édition nativement numérique
 - Recherche quantitative (notamment sur la variation orthographique)
- À l'origine, utilisation d'un lemmatiseur, puis passage à Excel pour la correction, puis retour au XML.

Premiers développements

Pandora Post-Correct App (ou PPA) qui va devenir Pyrrha dans les prochains mois. Open-Source pour le logiciel, aucune obligation quant aux données : <https://github.com/hipster-philology/pandora-postcorrect-app>

Objectifs :

1. Simplifier la correction sérielle : quand on corrige `Martins` vers le lemme `Martin` , **proposition** d'application de la correction à d'autres cas similaires
2. Historique des modifications
3. Listes de contrôle fournies par l'utilisateur-riche.

Exemple d'utilisation simple

1. L'utilisateur-riche donne des données généralement pré-lemmatisées suivant la forme

form	lemma	morph	POS
Ch'est	chëoir	NOMB.=s GENRE=m CAS=r	VERcjg
le	le	NOMB.=s GENRE=m CAS=r	DETdef
vie	vie1	NOMB.=s GENRE=f CAS=r	NOMcom
de	de	MORPH=empty	PRE
sainte	saint1	NOMB.=s GENRE=f CAS=r DEGRE=p	ADJqua
Baltelt	bataille	NOMB.=s GENRE=f CAS=r	NOMcom

Puis se rend sur une instance de type dev.chartes.psl.eu/ppa

Exemple d'utilisation simple : Création

Pyrrha Dashboard New Corpus Your Account Log out

Create a new corpus

Corpus Name
Enter name of the corpus
This should be a clear name

Left Context
3
Number of words to display on the left of the word to annotate

Right Context
3
Number of words to display on the right of the word to annotate

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

Load a configuration
If your language is part of the following option, we recommend using the default 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

Exemple d'utilisation simple : Modification

Pyrrha Dashboard							Your Account	Log out
et	et	CONcoo	MORPH=empty	hom fust saus et que cascuns venist	39	Save		
que	que4	CONsub	MORPH=empty	fust saus et que cascuns venist a	12	Save		
cascuns	chascun	PROind	NOMB.=s GENRE=m CAS=n	saus et que cascuns venist a le	1	Save		
venist	venir	VERcig	MODE=sub TEMPS=ipf PERS.=3 NOMB.=s	et que cascuns venist a le conaissance	0	Save		
a	a3	PRE	MORPH=empty	que cascuns venist a le conaissance de	13	Save		
le	le	DETdef	NOMB.=s GENRE=m CAS=r	cascuns venist a le conaissance de se	10	Save		
connaissance	conquestence	NOMcom	NOMB.=s GENRE=m CAS=r	venist a le connaissance de se verité	0	Save		
de	de	PRE	MORPH=empty	a le conaissance de se verité .	25	Save		
se	s	CONsub	PERS.=3 NOMB.=s GENRE=f CAS=r	le conaissance de se verité . Ses	1	Save		
verité	Sansuere! Sartaigne Satan Satanas Satanie Sodomien Sodomite	NOMcom	NOMB.=s GENRE=f CAS=r	connaissance de se verité . Ses nons	0	Save		
.		PONfrit	-	de se verité . Ses nons doit	13	Save		
Ses		DETpos	PERS.=3 NOMB.=p GENRE=m CAS=r	se verité . Ses nons doit estre	0	Save		

Exemple d'utilisation simple : Recherche d'incohérences

Pyrrha Dashboard New Corpus Issue62a Loherains Wauchier Essai 1 Martial 1.PR Bathilde Bathilde2 Your Account Log out

Quick links: Edit tokens Last Edit tokens

Corpus Bathilde2 - Similar tokens

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

All matches are at least a match on form.

1

Original token

Form	Context	Lemma	POS	Morph
cascuns	qui voudroit que cascuns hom fust saus	chascun	DETind	NOMB.=s GENRE=m CAS=n

Similar matching

Form	Lemma	POS	Morph	Context	Save
cascuns	chascun	PROind	NOMB.=s GENRE=m CAS=n	saus et que cascuns venist a le	Save

0

Exemple d'utilisation simple : Historique

Corpus Loherains - List of tokens

1 2 3 4 5 ... 353 354

User	Edit	Form	Context	Lemma	POS	Morph	Corr Lemma	Corr POS	Corr Morph	Similar	Actions
A.Cochet Cardillo	2018-11-01 22:26:44	aceri	oixel chantent doucement aceri De l'autre part	acerin	OUT	NOMB.=p GENRE=m CAS=n	asserir	VERppe	NOMB.=s GENRE=m CAS=r	0	Find Similar
A.Cochet Cardillo	2018-11-01 22:22:06	Lassus	aigue fist vin Lassus en cel palais	là	ADVgen	DEGRE=-	la_sus	ADVgen	DEGRE=-	0	Find Similar
A.Cochet Cardillo	2018-11-01 22:19:14	guier	penst Dieus del guier De Saint Denis	guier	VERinf	MORPH=empty	guier	NOMcom	NOMB.=s GENRE=m CAS=r	0	Find Similar
A.Cochet Cardillo	2018-11-01 22:18:44	del	or penst Dieus del guier De Saint	de	OUT	NOMB.=s GENRE=m CAS=r	de+He	PRE.DETdef	NOMB.=s GENRE=m CAS=r	0	Find Similar
A.Cochet Cardillo	2018-11-01 22:17:54	Hervi.	devomes dou duc Hervi . parler Ki chevaucha	Hervis	NOMpro	NOMB.=s GENRE=m CAS=r	SEGM	OUT	NOMB.=s GENRE=m CAS=r	0	Find Similar
A.Cochet Cardillo	2018-11-01 22:17:38	fort.	biaus freires au fort . roi me direz	fort	ADJqua	NOMB.=s GENRE=m CAS=r DEGRE=p	SEGM	OUT	NOMB.=s GENRE=m CAS=r DEGRE=p	0	Find Similar

Deucalion

<http://github.com/hipster-philology/deucalion>

Problématiques

- Lemmatiseur = travail important de développement informatique ET d'annotation/travail philologique.
- Lemmatiseurs ont des interfaces différentes

Objectifs non-couverts

Deucalion **n'est pas** un service de lemmatisation massive. Il est prévu principalement pour une utilisation en lien avec Pyrrha/PPA afin d'accélérer le travail de l'annotateur-riche.

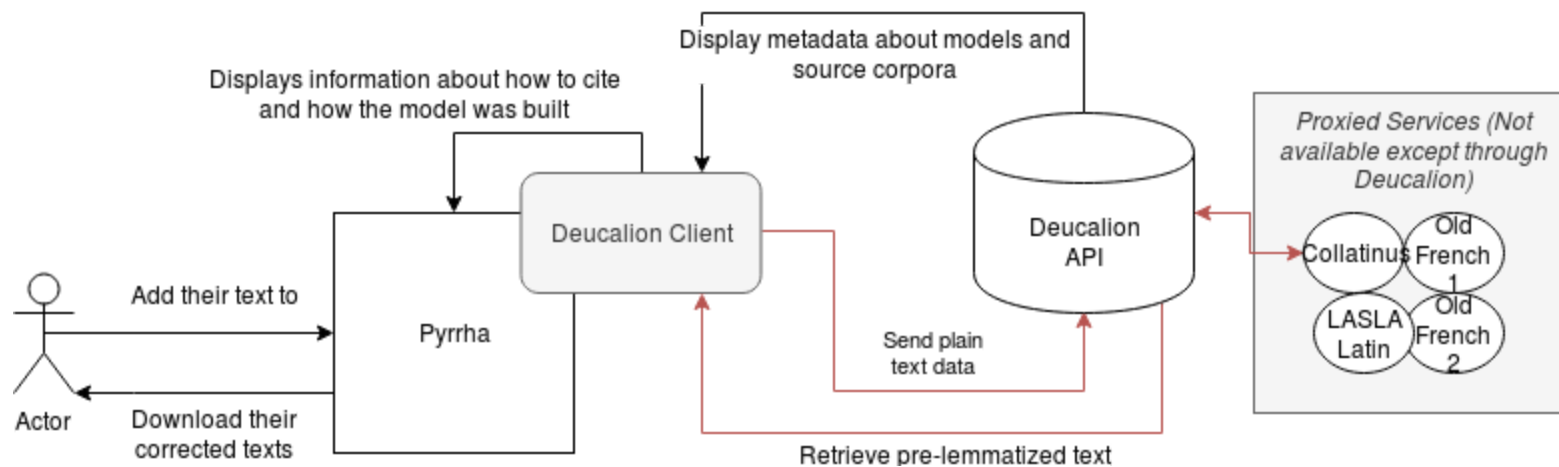
Deucalion :Modèle actuel (non définitif)

```
"@id":"http://127.0.0.1:5005/model/enc-001",
"dc:creator":[{"@id":"https://viaf.org/viaf/134649805/",
"@label":"Ecole nationale des Chartes"}],
"dc:description":["Model built on Wauchier de Denain, Graal
and Old French data"],
"dc:language":"fro",
"dc:source":[{"dc:alternative":"Ariane Pinche, \"Li Seint Confessors, Édition
nativement numérique\". 2018-10-23, Available at
http://chartes.psl.eu/corpora/Wauchier",
"dc:authors":[{"@id":"http://chartes.psl.eu/apinche",
"@label":"Ariane Pinche"}],
"dc:date":"2018-10-23",
"dc:title":"Li Seint Confessors, \u00c9dition nativement num\u00e9rique"}],
"dc:title":"Model for Ancient-French"}]
```

Ce qu'il manque

- Une description fine du logiciel utilisé. Une propriété générale de citation (comme `dc:alternative` pour les Sources ?)
- Un client, afin de montrer automatiquement à l'utilisateur-riche les modèles et outils disponibles, mais aussi comment les citer (contrairement à la situation actuelle "Ancien Français").
- Information sur les financeurs (moins important pour certains pays que d'autres.)

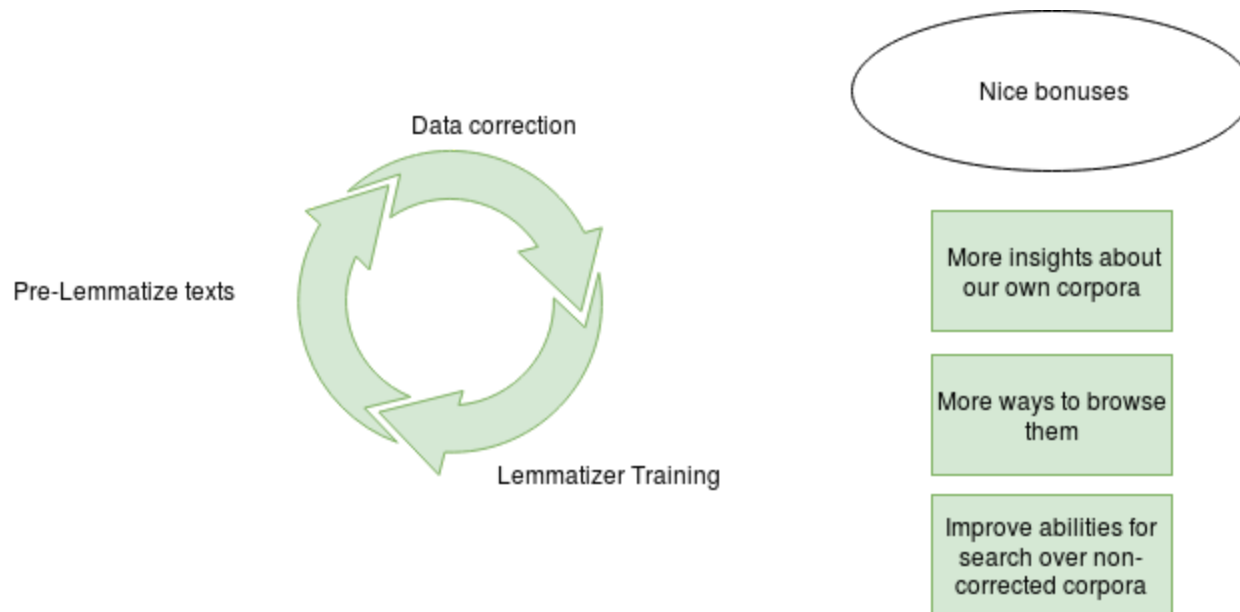
L'architecture



À Implémenter

- Amélioration des performance de l'application [Issue 64](#)
- *Inter-Corrector Agreement* [Issue 61](#)
- Afficher un identifiant ou numéro d'ordre des mots [Issue 69](#)
- Ajout de colonnes au choix de l'utilisateur-riche [Issue 13](#)
- Consistency Check POS / Morph [Issue 25](#)
- Statut vérifié [Issue 29](#)

En définitif



Merci

Liens :

- <http://github.com/hipster-philology/deucalion>
- <https://github.com/hipster-philology/pandora-postcorrect-app>
- <https://github.com/PonteIneptique/deucalion-model-af>