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-rice.

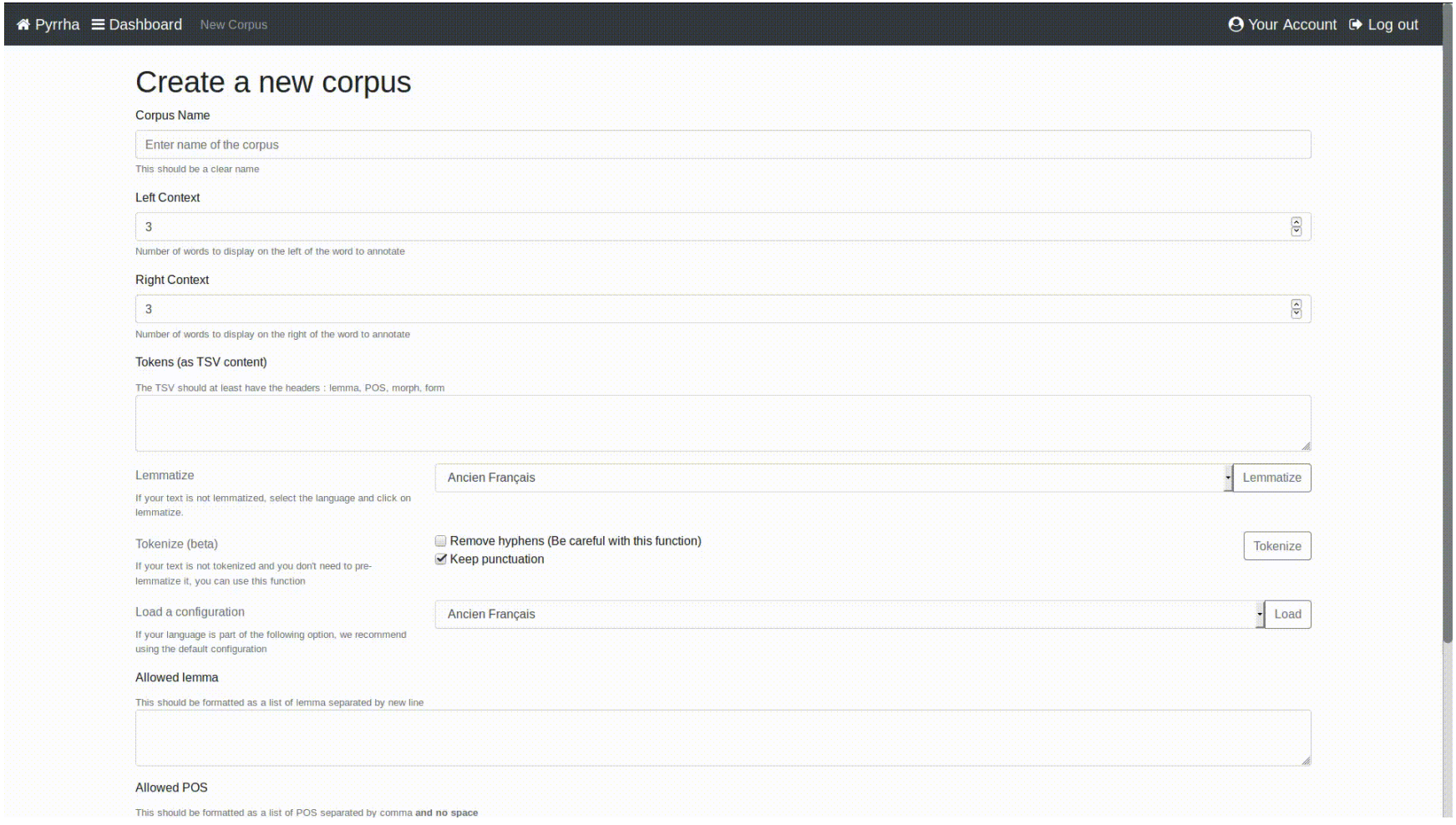
Exemple d'utilisation simple

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

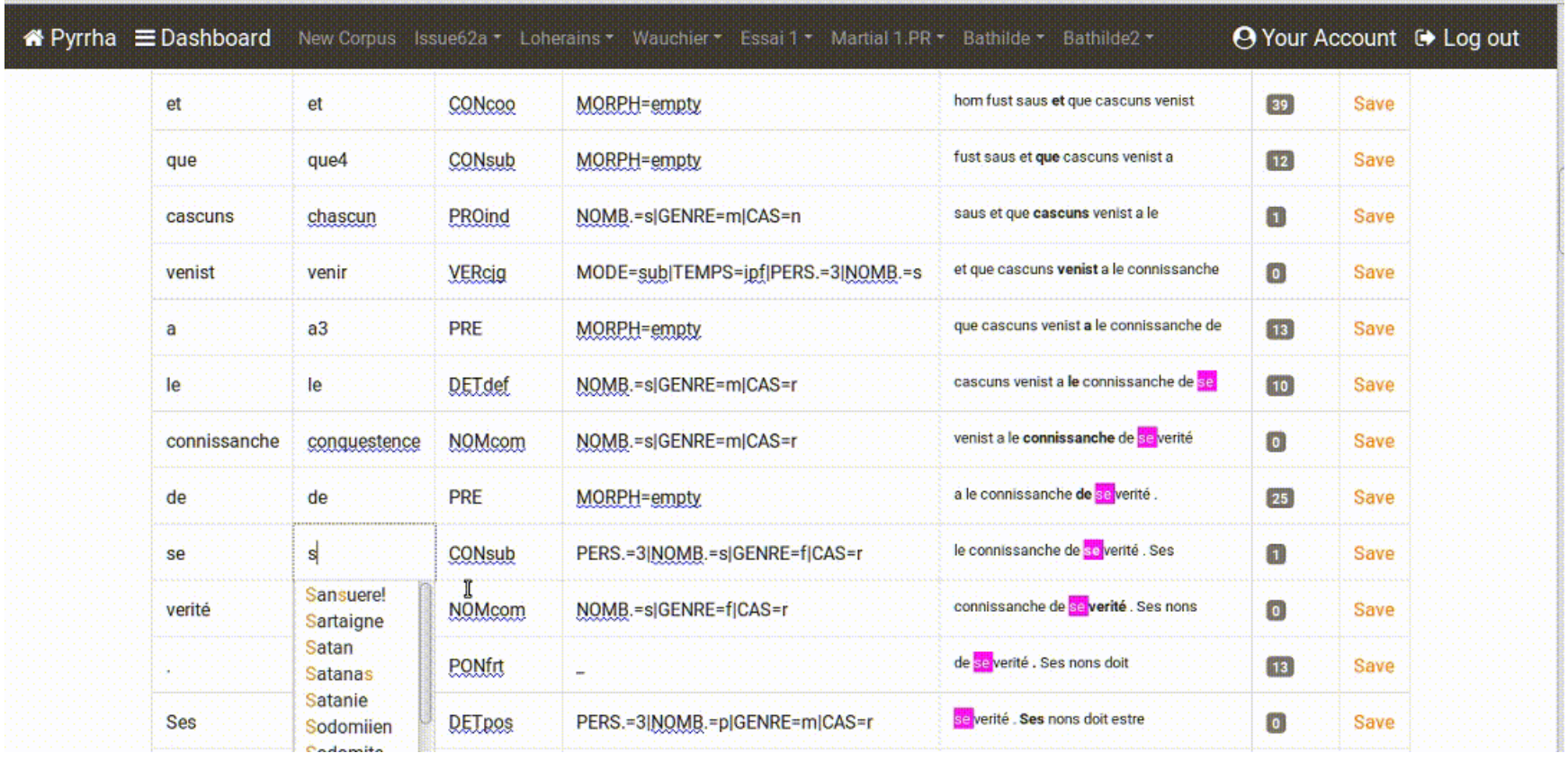
```
form    lemma   morph   POS
Ch’est  chëoir  NOMB.=s|GENRE=m|CAS=r   VERcjg
le      le      NOMB.=s|GENRE=m|CAS=r   DETdef
vie     vie1    NOMB.=s|GENRE=f|CAS=r   NOMcom
de      de      MORPH=empty     PRE
sainte  saint1  NOMB.=s|GENRE=f|CAS=r|DEGRE=p   ADJqua
Baltelt bataille        NOMB.=s|GENRE=f|CAS=r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

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

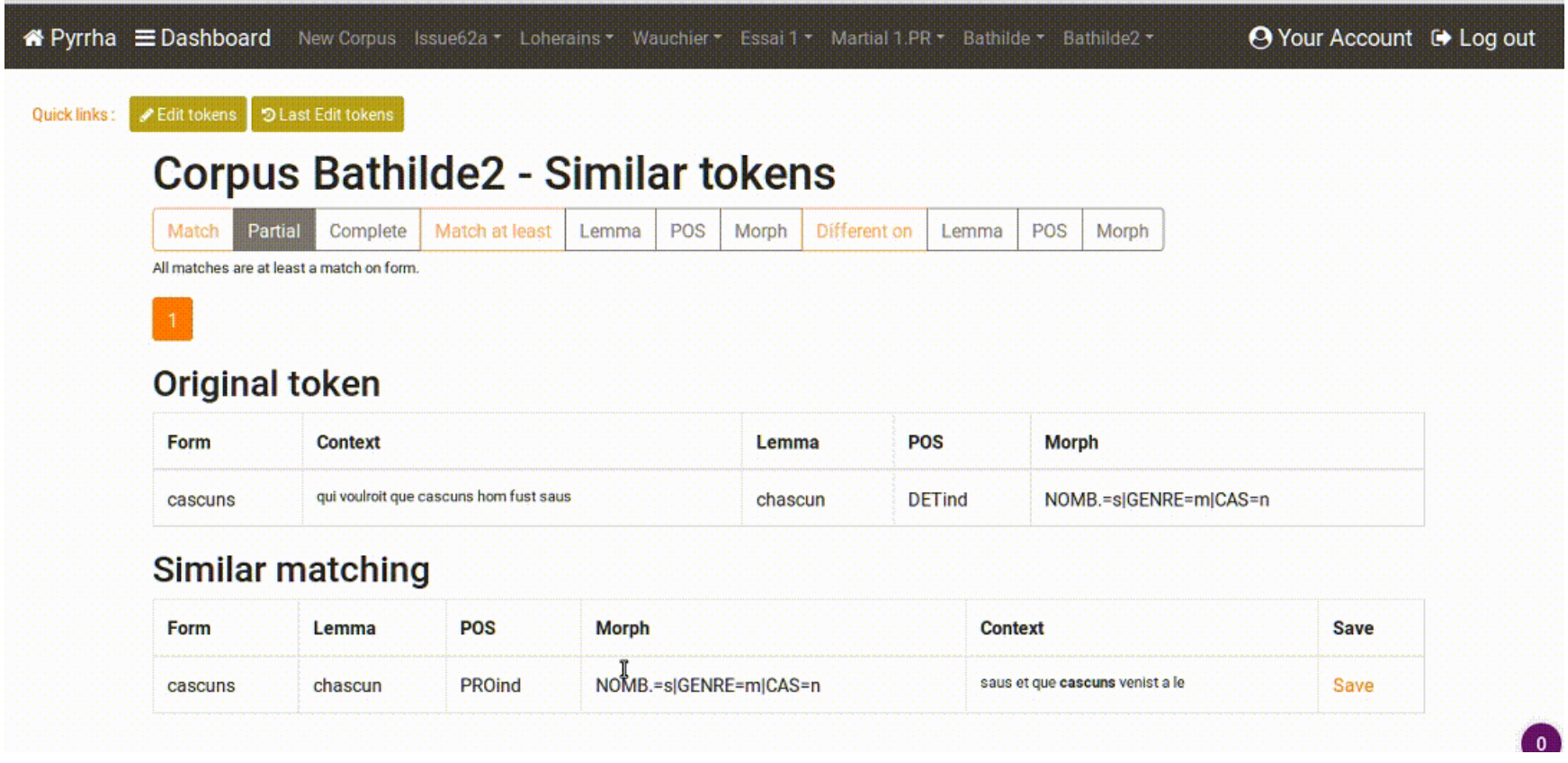
Exemple d'utilisation simple : Création



Exemple d'utilisation simple : Modification



Exemple d'utilisation simple : Recherche d'incohérences



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	guiier	VERinf	MORPH=empty	guiier	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+le	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'accélerer le travail de l'annotateur-rice.

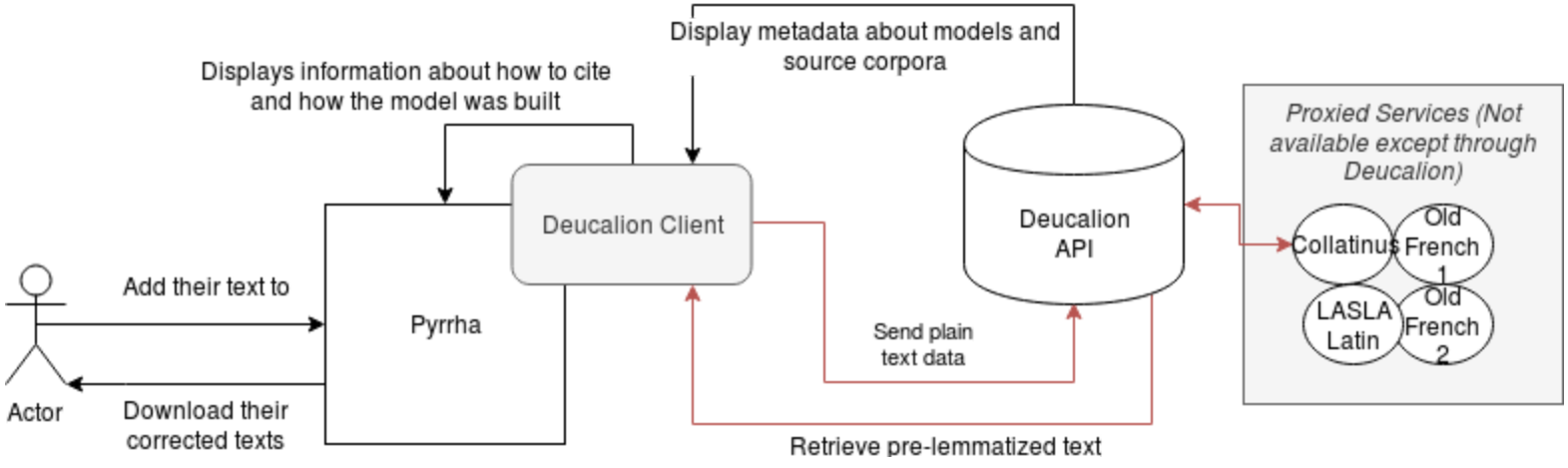
Deucalion :Modèle actuel (non définitif)


```
"@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\u00e
   }
],
"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-rice 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.)

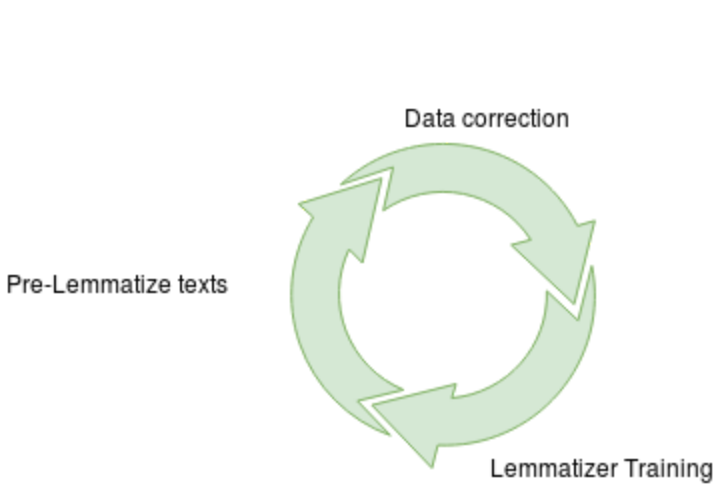
L'architecture



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-rice Issue 13
- Consistency Check POS / Morph Issue 25
- Statut vérifié Issue 29

En définitif



Nice bonuses

More insights about our own corpora

More ways to browse them

Improve abilities for search over non-corrected corpora

Merci

Liens :

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