

18 DECEMBRE 2023

PROJET FINAL

SEMANTIQUE COMPUTATIONNELLE

PICARD CLARA SORBONNE UNIVERSITE



Table des matières :

- I. Introduction
- II. Création de corpus et du DataStore
- III. Annotations manuelles
- IV. Annotations avec dictionnaire
- V. Annotations automatiques
- VI. Impressions des annotations en html
- VII. Détection des langues

Conclusion



I. Introduction

Pour ce projet, nous avons utilisé le logiciel GATE pour faire différents types d'annotations. Nous avons choisi le livre Pride and Prejudice de Jane Austen en anglais et nous avons sélectionné la préface ainsi que les onze premiers chapitres, ce qui fait plus de 15 000 mots. Nous avons décidé de prendre ce livre parce qu'il y a beaucoup de personnages, lieux et dialogues et donc une grande variété de verbes de discours et de mouvements.

Nous avons téléchargé notre livre sur la plateforme du *Projet Gutenberg* qui propose tous les livres libres de droits dans plusieurs langues dont l'anglais et le français. Pour notre projet, nous avons téléchargé les fichier en format Plain Text UTF-8 pour faciliter la division des chapitres. Nous avons mis chaque chapitres dans des fichier .txt que nous avons mis dans un dossier nommé *Livres sélectionnés*, qui nous servira pour créer le corpus dans GATE.



II. Création du corpus et du DataStore

Une fois le dossier conçut avec tous les chapitres, nous avons créé un corpus contenant notre livre dans GATE en faisant un clic droit sur *Language Resources* puis en cliquant sur *New* et finalement sur *GATE Corpus*, nous avons nommé le corpus *Corpus Anglais*. Il a ensuite fallu remplir ce corpus avec nos chapitres. Pour se faire, il faut faire un clic droit sur *Corpus Anglais* et sélectionner *Populate* pour choisir le dossier contenant notre livre.

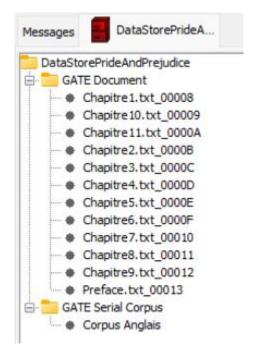


Nous avons ensuite créé un Datastore pour ce corpus, cela permet de pouvoir enregistrer nos annotations et de les retrouver. Pour créer ce Datastore il faut tout d'abord créer un dossier nommé *DataStorePrideandPrejudice* puis sur GATE faire un clic droit sur *Datastores* et cliquer



sur *Create Datastore* et sur *SerialDataStore*, nous avons ensuite sélectionné le dossier créé précédemment et voici notre datastore créé.





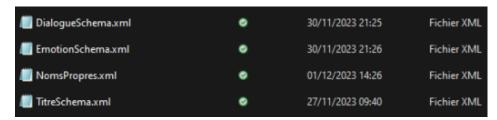
III. Annotations manuelles

Dans un second temps, nous avons réalisé des annotations manuelles en utilisant les schémas d'annotations et les expressions régulières. Le première étape est de créer des schémas d'annotations pour annoter les titres et les dialogues. Pour cela, il faut dans un bloc note comme Notepad++ créer nos schémas d'annotations que nous avons enregistré au format .xml. Nous avons repris les schémas d'annotations fait en cours et nous les avons modifiés avec nos paramètres. Nous avons décidé d'annoter les titres (des livres et des chapitres), les noms propres (prénom, nom, famille, lieu, titre et jour), les dialogues (les monologues et dialogues) ainsi que les émotions (positives et négatives).

```
<?xml version="1.0"?>
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
   <!-- XSchema definition for Person -->
   <element name="Titre">
      <complexType>
       <attribute name="kind" use="required">
          <simpleType>
            <restriction base="string">
              <enumeration value="livre"/>
              <enumeration value="chapitre"/>
            </restriction>
          </simpleType>
        </attribute>
      </complexType>
    </element>
 </schema>
```

```
?xml version="1.0"?
<schema xmlns="http://www.w3.org/2000/10/XMLSchema">
  <!-- XSchema definition for Person
  <element name="Noms Propres">
    <complexType>
      <attribute name="kind" use="required">
        <simpleType>
          <restriction base="string">
            <enumeration value="Prénom"/>
            <enumeration value="Nom"/>
            <enumeration value="Famille"/>
            <enumeration value="Lieu"/>
            <enumeration value="Titre"/>
            <enumeration value="Jour"/>
          </restriction>
        </simpleType>
      </attribute
    </complexType>
  </element>
</schema>
```

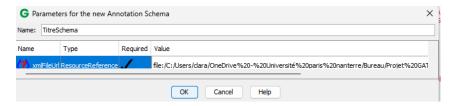




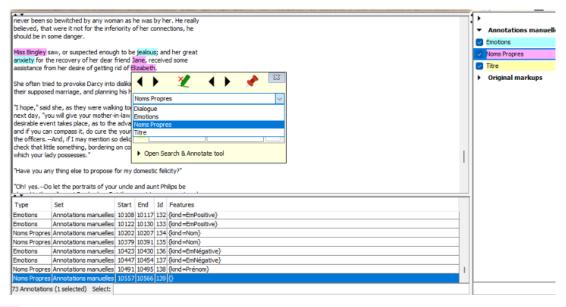
Une fois les schémas d'annotations créés, pour les ajouter dans GATE il faut aller dans Language Resources, cliquer sur New, puis Annotation Schema:

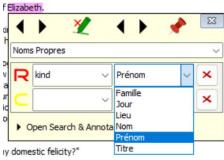


L'étape suivante est de sélectionner les schémas d'annotations créés précédemment :



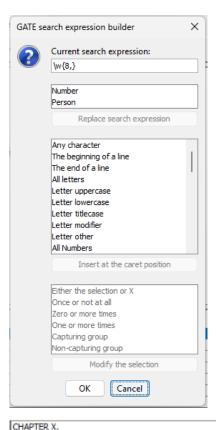
Et ensuite sélectionner les mots/expressions que nous souhaitons annoter et de choisir comment l'annoter (Titre, Noms Propres, Emotion ou Dialogue) :







Pour faire des annotations manuelles avec des expressions régulières, il faut sélectionner dans le texte un mot que nous souhaitons annoter, dans la page qui s'ouvre cocher la case *Regexp* et aller sur le point d'interrogation pour entrer l'expression régulière que nous avons choisis. Nous avons décidé d'annoter les mots longs et nous avons estimé qu'un mot long est un mot qui a plus de 8 lettres. Voici mon expression régulière : \w{8,}. Il faut ensuite cliquer su *First* pour qu'il trouve la première occurrence puis sur *Next* et enfin sur *Ann. all next* pour qu'il annote tous les mots du texte de plus de 8 lettres. Nous avons mis ces annotations dans l'ensemble d'annotation *Annotations manuelles*.





The day passed much as the day before had done, Mrs. Hurst and Miss Bingley had spent some hours of the morning with the invalid, who continued, though slowly, to mend; and in the evening Elizabeth joined their party in the drawing-room. The loo table, however, did not appear. Mr. Darcy was writing, and Miss Bingley, seated near him, was watching the progress of his letter, and repeatedly calling off his attention by messages to his sister. Mr. Hurst and Mr. Bingley were at piquet, and Mrs. Hurst was observing their game. Elizabeth took up some needlework, and was sufficiently amused in attending to what passed between Darcy and his companion. The perpetual commendations of the lady either on his hand-writing, or on the evenness of his lines, or on the length of his letter, with the perfect unconcern with which her praises were received, formed a curious dialogue, and was exactly in unison with her opinion of each. "How delighted Miss Darcy will be to receive such a letter!" He made no answer. "You write uncommonly fast." "You are mistaken. I write rather slowly." "How many letters you must have occasion to write in the course of the year! Letters of business too! How odious I should think them! "It is fortunate, then, that they fall to my lot instead of to yours."

| | Dialogue |
|----|--------------|
| | Emotions |
| V) | MotLong |
| | Noms Propres |
| | Titre |



IV. Annotations avec dictionnaire

Nous avons ensuite créé quatre dictionnaires pour annoter les personnages, les lieux, les événements et les références temporelles. Pour ce faire, nous avons dressé des listes avec les personnages du livres, les lieux, les événements et les références temporelles. Nous avons décidé de faire quatre listes pour les personnages : la famille Bennet, la famille Bingley, la famille Darcy et les personnages secondaires. Pour les lieux, nous avons aussi fait quatre listes : les noms des maisons des personnages, les villes, les régions et les pays. Pour les événements, nous avons décidé d'annoter les jours, les mois et les dates complètes/années. Et pour les références temporelles nous avons simplement dressé une listes de mots/expressions avec par exemple « this morning », « while », « later », etc. Nous avons ensuite créé les définitions de ces listes :

Exemple liste personnages:

John Willoughby George Wickham Wickham Fitzwilliam Edward Gardiner Lady Lucas Charlotte Lucas George Knightley Knightley Frederick Wentworth Wentworth William Lucas William Colonel Forster Mr. Philips Captain Carter Mrs. Hurst Mr. Hurst

Définition personnages :

bennet.lst:Personnage:FamBennet:Eng:FamBennet
bingley.lst:Personnage:FamBingley:Eng:FamBingley
darcy.lst:Personnage:FamDarcy:Eng:FamDarcy
persosecondaire.lst:Personnage:PersoSecondaire:Eng:PersoSecondaire

Définition lieux :

ville.lst:Lieu:Ville::Ville
pays.lst:Lieu:Pays::Pays
region.lst:Lieu:Region::Region
maison.lst:Lieu:Maison::Maison

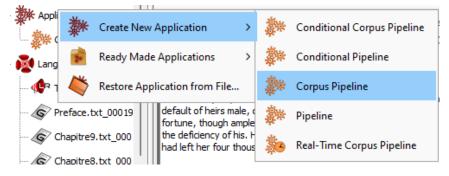
Définition évènements :

evenements.lst:Evenement::Eng:Date jour.lst:Evenement::Eng:Jour mois.lst:Evenement::Eng:Mois

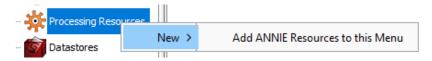
Définition références temporelles :

reftempeng.lst:RefTemporelle::Eng:RefTemporelle

Pour faire les annotations dans GATE, il faut dans un premier temps créer un Corpus Pipeline en allant dans *Applications*, puis *Create New Application* et *Corpus Pipeline* :

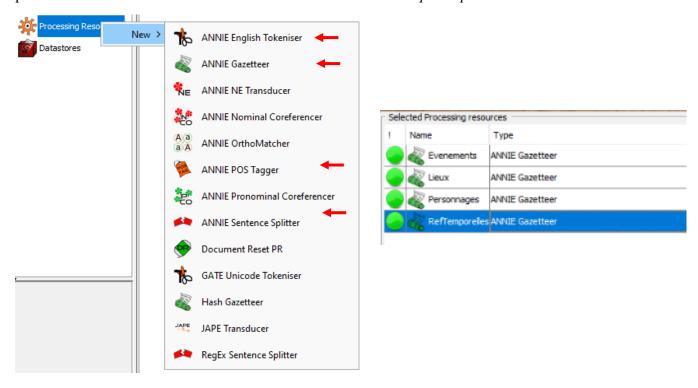


Ensuite, il faut aller dans *Processing Resources* et ajouter *ANNIE Resources* au Menu :





Dans ce Menu, nous avons sélectionner ANNIE English Tokeniser, ANNIE Sentence Splitter, ANNIE POS Tagger et ANNIE Gazetteer dans lequel nous ajoutons nos définitions créées précédemment et nous avons mis tous ses éléments dans le Corpus Pipeline.



CHAPTER VII.

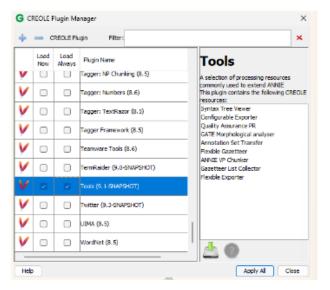
Voici quelques résultats de nos annotations avec les dictionnaires :



Bennet's property consisted almost entirely in an estate of two thousand a year, which, unfortunately for his daughters, was entailed in default of heirs male, on a distant relation; and their mother's fortune, though ample for her situation in life, could but ill supply the deficiency of his. Her father had been an attorney in M had left her four thousand pounds. She had a sister married to a Mr. Philips, who had been a derk to their father, and succeeded him in the business, and a brother settled in London in a respectable line of trade. The village of Longbourn was only one mile from Meryton; a most convenient distance for the young ladies, who were usually tempted thither three or four times a week, to pay their duty to their aunt and to a miliner's shop just over the way. The two youngest of the family, Catherine and Lydia, were particularly frequent in these attentions; their minds were more vacant than their sisters', and when nothing better offered, a walk to Meryton was necessary to amuse their morning

Après cette étape, nous avons choisis d'annoter les verbes de mouvements et de discours du corpus à l'aide de dictionnaires flexibles. Pour cela, il faut utiliser un Hash Gazetteer, un Flexible Gazetteer et un Morphological Analyser que nous avons activer dans *Creole Plugin* avec *Tools*:



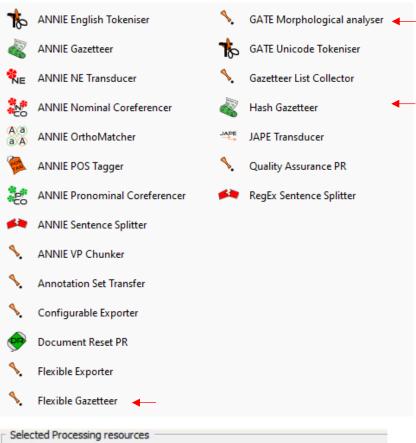


Nous avons ensuite dressé des listes pour les verbes de mouvements et de discours à l'infinitif et nous avons créé leurs définitions :



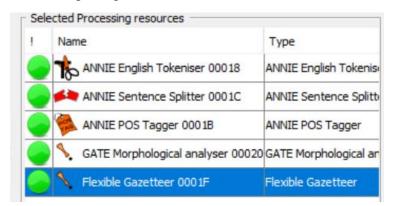
Dans GATE, il faut mettre les définitions dans des *Hash Gazetteer* et les exécuter dans un premier temps avec un *Morphological analyser* pour annoter les verbes à l'infinitif :



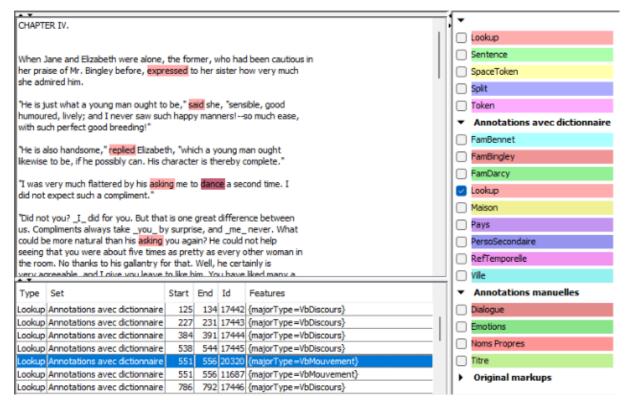


| | Name | Туре |
|---|-----------------------------------|-----------------------------|
| 6 | ANNIE English Tokeniser 00025 | ANNIE English Tokeniser |
| | ANNIE Sentence Splitter 00029 | ANNIE Sentence Splitter |
| | ANNIE POS Tagger 00028 | ANNIE POS Tagger |
| | SATE Morphological analyser 0002D | GATE Morphological analyser |
| | Mouvement | Hash Gazetteer |

Une fois l'application exécutée, nous avons vérifié que les verbes étaient bien annotés puis nous avons enlevé les *Hash* Gazetteer du *Corpus Pipeline* pour les remplacer par le *Flexible Gazetteer* pour qu'il annote toutes les flexions des verbes.







V. Annotations automatiques

Pour les annotations automatiques avec Jape, nous avons décidé d'annoter les moments importants du livre, c'est-à-dire quand une phrase commence par « when », « after », « but », etc. Dans ces moments importants, nous avons annoté les personnages, les verbes et le contexte/moment. Pour cela, nous avons créé des grammaires Jape. Pour la grammaire des moments importants, nous avons mis les adverbes, les pronoms et les prépositions, etc., puis nous avons fait en sorte que l'annotation se finisse par un point. Pour les personnages, nous avons dressé une liste des personnage dans un fichier .lst et nous avons créé sa définition puis dans la grammaire Jape, nous avons appelé la liste des personnages :

```
Phase: Personnage
Input: Token SpaceToken Phrase
Options: control = appelt

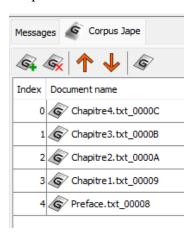
Rule: ReglePersonnage
Priority: 40
(
{Phrase contains {Token.category == Personnage}})
):personnage
-->
:personnage.Personnage = {}
```

Dans la grammaire des verbes, nous avons mis tous les verbes (VB, VBD, VBN, VBG VBP). Et enfin, pour la grammaire des contextes/moments nous avons décidé d'annoter à partir des adverbes, pronoms et prépositions jusqu'à une virgules.

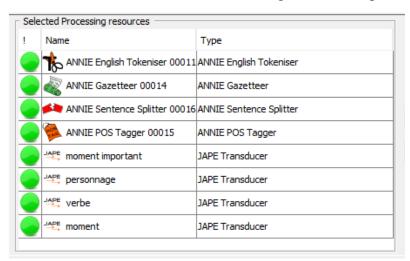
Pour annoter dans GATE, il faut ajouter nos grammaires Jape dans des *Jape Transducer* et les exécuter dans le *Corpus Pipeline* avec le *ANNIE Gazetteer* des Personnage. Nous avons

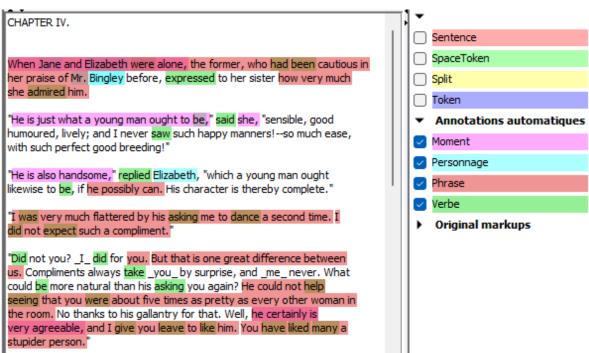


rencontré un problème d'exécution, quand nous cliquions sur *Run this application*, le logiciel GATE ne répondais plus. Nous avons donc décidé de créer un nouveau datastore avec moins de chapitres :



Une fois cela fait, nous avons exécuté les grammaires Jape sur notre Corpus Jape :



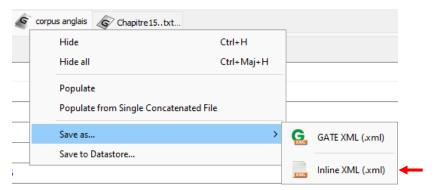




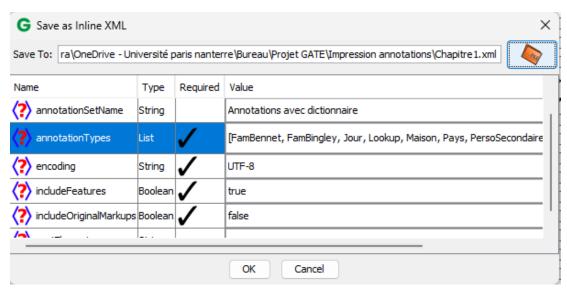
Nous pouvons voir que nos grammaires Jape ne sont pas assez optimisées parce qu'elles annotent aussi les verbes, personnages et contextes en dehors des moments importants.

VI. Impressions des annotations en html

Pour l'impressions des annotations en HTML, il faut dans un premier temps enregistrer le corpus en XML :



Puis dire où enregistrer le fichier XML et donner l'ensemble d'annotations que nous voulons imprimer (*annotationSetName*) et les annotations (*annotationTypes*) :



Il faut ensuite ouvrir le fichier XML dans Notepad++ et ajouter en haut du fichier les balises HTML et mettre le corps du texte dans une balise

 dody> :

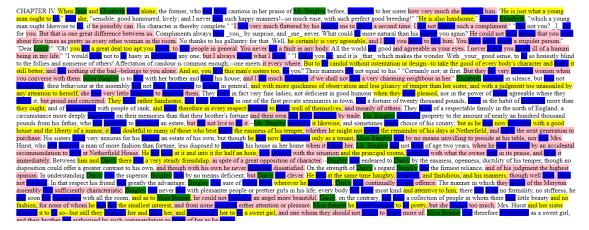


Il faut maintenant créer une feuille CSS. Pour cela, il faut dans notre fichier créé précédemment ajouter dans <head> les couleurs CSS. Nous avons choisi d'annoter les phrases en jaune, les personnages en rouge, les verbes en bleu (mais cela peut changer en fonction des annotations que nous voulons imprimer, comme sur la deuxième photo pour le chapitre 2). Pour cela, il faut utiliser « background-color: ». Une fois cela fait, nous avons enregistrer le fichier en .html.

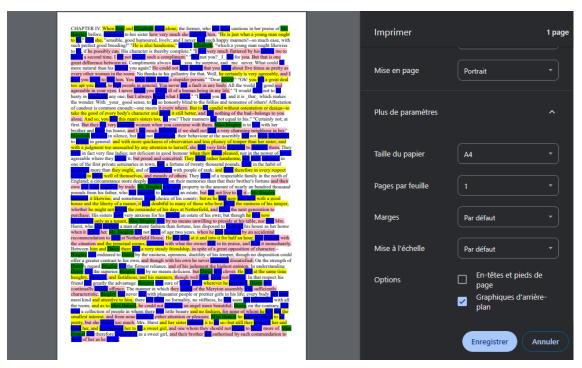
```
meta charset="utf-8"
                             <link rel="stylesheet" href=""exemple.css";</pre>
                                      Verbe {
                                                background-color: blue:
                                                 background-color: pink;
                                                background-color: yellow
                                                background-color: green;
                             </style
                           <Moment gate:gateId="3283">When <Personnage gate:gateId="3203" majorType="Personnage" language="" minorType="">Jane</Personnage> and <Personnage gate:gateId="3207" majorType="personnage" language="" minorType="">Mr. <Personnage gate:gateId="3209" majorType="Personnage" language="" minorType="" minorType=" minorTy
                           squot/<Moment gate:gateId="3286">He is also handsome,</Moment>squot; <Verbe gate:gateId="3366">replied</Verbe> <Personnage gate:gateId="3211" majorType="Personnage" language=
likewise to <Verbe gate:gateId="3367">be</Verbe>, if <Fhrase gate:gateId="3319">he possibly can.</Fhrase> His character is thereby complete.squot;
                             squot;<Phrase gate:gateId="3320">I <Verbe gate:gateId="3368">was</Verbe> very much flattered by his <Verbe gate:gateId="3369">asking</Verbe> me to <Verbe gate:gateId="3370">d
<!DOCTYPE html>
                             <meta charset="utf-8>
                            <link rel="stylesheet" href=""exemple.css">
                                         FamBennet {
                                          FamBingley {
                                                      background-color: pink;
                                         Jour (
                                                      background-color: grey;
                                         Lookup {
                                                      background-color: red;
                                                      background-color: vellow:
                                         Pays (
                                                      background-color: green;
                                          PersoSecondaire (
                                                      background-color: purple:
                                         RefTemporelle (
                                                     background-color: orange;
                             </style>
               c/head
                             CHAPTER II.
```

CHAPTER II. M. Bennet was among the earliest of those who waited on Mr. Bingley. He had always intended to visit him, though to the last always assume his wife that he should not real after the visit was paid, she had no knowledge of it. It was there disclosed in the following manner. Observing his second daughter employed in trimming a hat, he suddenly addressed her with, "I hope Mr. Bingley will like it Lizzy." "We are not in a way to know _what_Mr. Bingley likes," as her mother resentfully, "innee we are not to visit." "But you forget, mama," and introduce him." I'd on to believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her." "No more have I," and Mrs. Long has promised to introduce him." I'd on not believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her." "No more have I," and Mrs. Long has promised to introduce him." I'd on the believe Mrs. Long will do any such thing. She has two nieces of her own. She is a selfish, hypocritical woman, and I have no opinion of her." "No more have I," and Mrs. Long has promised to introduce him. For sake! Have a little compassion on my nerveys. You tear them to pieces." "Kim has no discretion in her sum?" I will be a sake! Have a little compassion on my nerveys. You tear them to pieces." "Kim has no discretion in her sum?" I will be a sake! Have a little compassion on my nerveys. You tear them to pieces." "Kim has no discretion in her sum?" I will be a sake! Have a little compassion on my nerveys. You tear them to pieces." "Kim has no discretion in her sum?" I will be for the introduce him, for she will not know him herself." "The girls of the compassion on the pieces." "The man of the contain herself, began in the sake and th



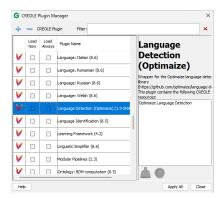


Nous avons visualisé notre page HTML dans notre navigateur puis nous l'avons enregistré au format pdf :



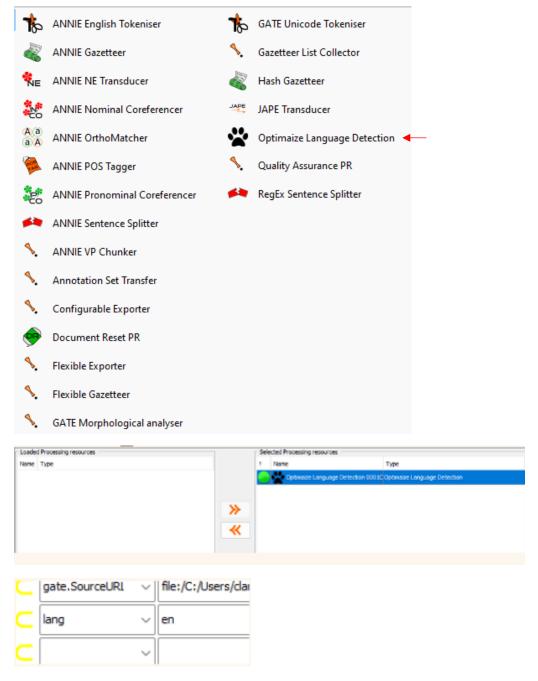
VII. Détection des langues

Pour faire la détection des langues, il faut aller dans CREOLE Plugin Manager, chercher Language Detection et l'appliquer :





Dans *Processing Resources*, il faut choisir *Language Detection (Optimaize)* et aller dans le *Corpus Pipeline* pour appliquer l'application :



Nous pouvons voir que l'application a bien détecté que la langue du corpus est l'anglais.

VIII. Conclusion

Pour conclure, nous avons vu différentes manières d'annoter des documents sur GATE. Selon nous, ces méthodes présentent des avantages comme des inconvénients. Avec des annotations manuelles les annotations sont plus précises mais demandent plus de temps que les annotations avec dictionnaire ou les annotations automatiques. Les annotations avec dictionnaires sont moins précises dans certains cas comme par exemple pour le mois de mai qui va être annoter « May » pour les mois mais aussi « may » pour le verbe si on enlève la sensibilité à la casse.



Les annotations automatiques avec Jape sont elles aussi moins précises que les annotations manuelles parce que si elles ne sont pas bien faites, elles vont annoter des éléments que nous ne désirons pas annoter. L'avantage des annotations avec dictionnaire et automatiques est que c'est beaucoup plus rapide. Le choix dépend donc des besoins spécifiques du projet.