Projet final

Présenté par

Chengwanli YANG

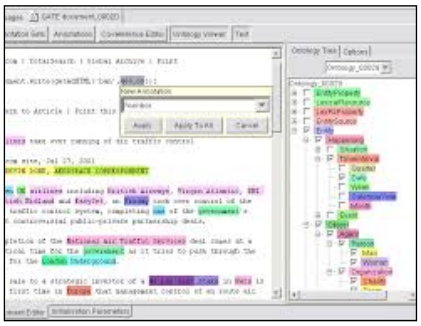
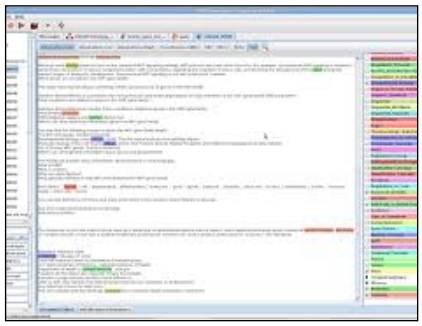
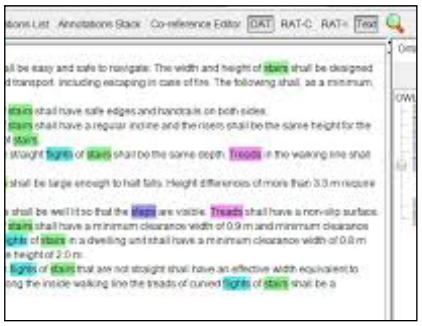
Master 1

Sémantique computationnelle

Sorbonne Université —Faculté des Lettres
Décembre 2020

Introduction

GATE Developer est un environnement de développement qui fournit un riche ensemble d'outils graphiques interactifs pour la création, la mesure et la maintenance de composants logiciels du traitement de langage humain.

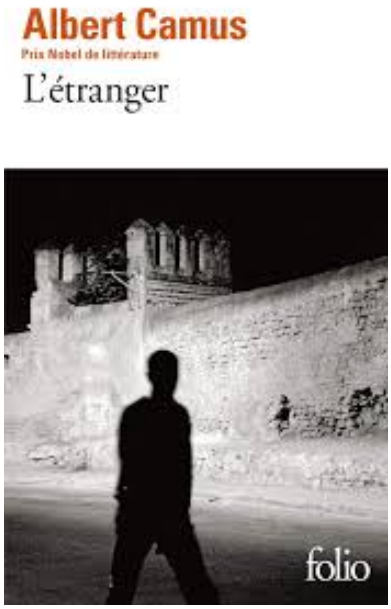
Objectif

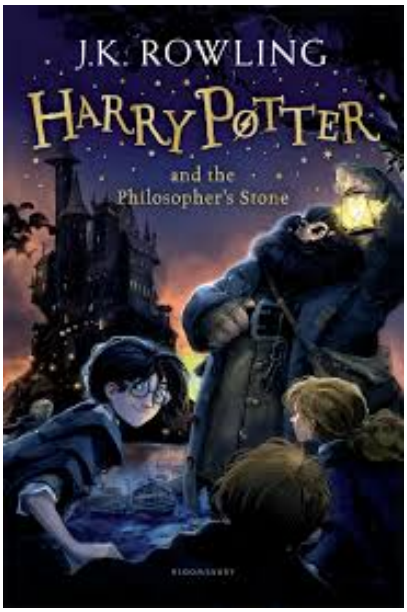
L'objectif est d'annoter manuellement et automatiquement deux livres, un en français et un en anglais en utilisant GATE 8.6.1.

Description

1. Sélection des livres et création des corpus:

Je choisis *L'étranger* d'Albert Camus et *Harry Potter and the Philosopher's Stone* de J.K. Rowling comme les corpus, en traitant les 5 premiers chapitres.



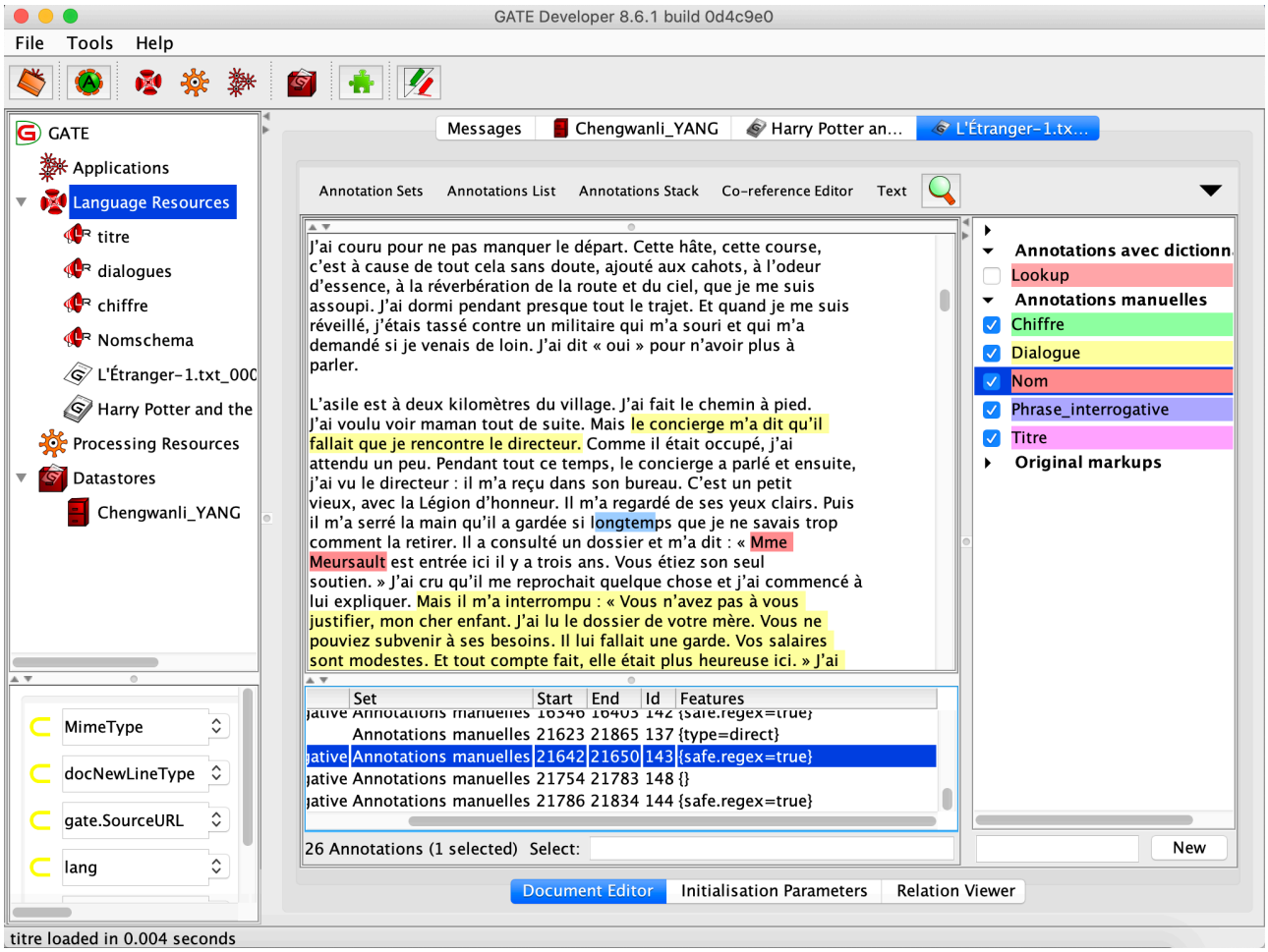


2. Création d'un data store :

La data store est nommé « Chengwanli_YANG ».

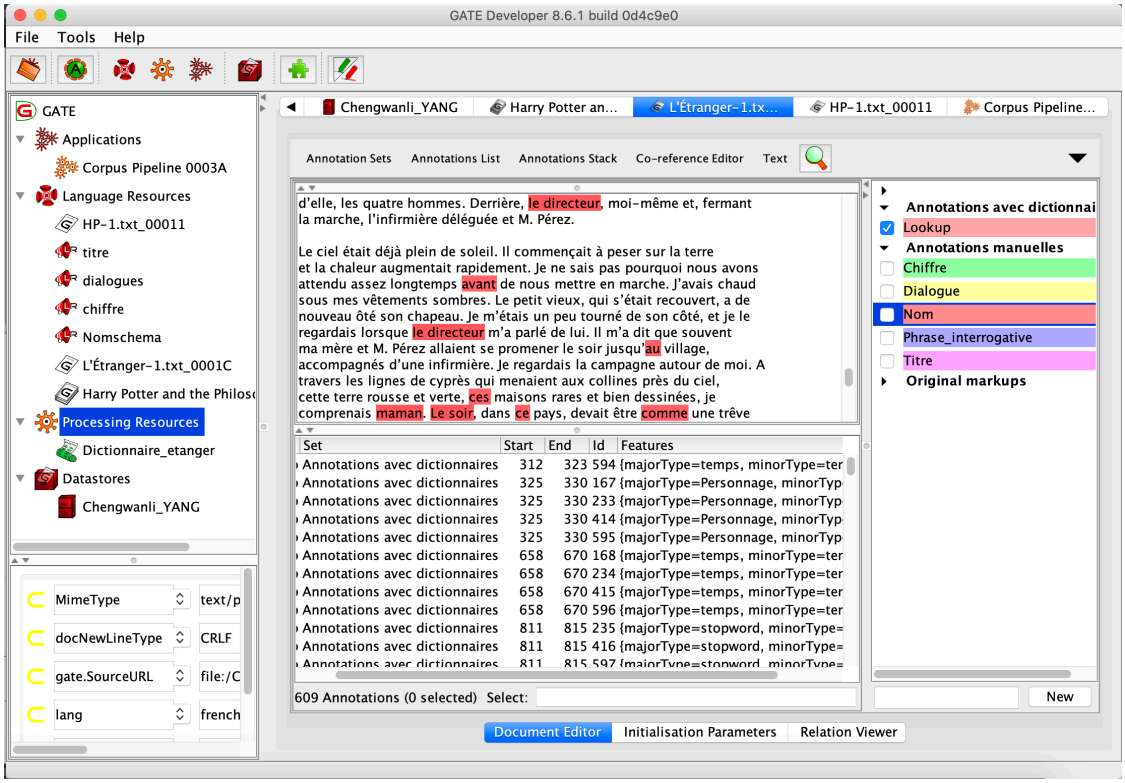
3. Annotations manuelles :

Dans le document « L'Étranger-1.txt_0001C » du corpus « Étranger », je crée d'abord 4 schémas « Chiffre », « Nom », « Titre », « Dialogues ». Dans le schéma « Titre », je précise l'autre, l'éditeur, le chapitre et l'année d'édition etc. De plus, je fais les annotations nommées « Phrase_interrogative » avec l'expression régulière.

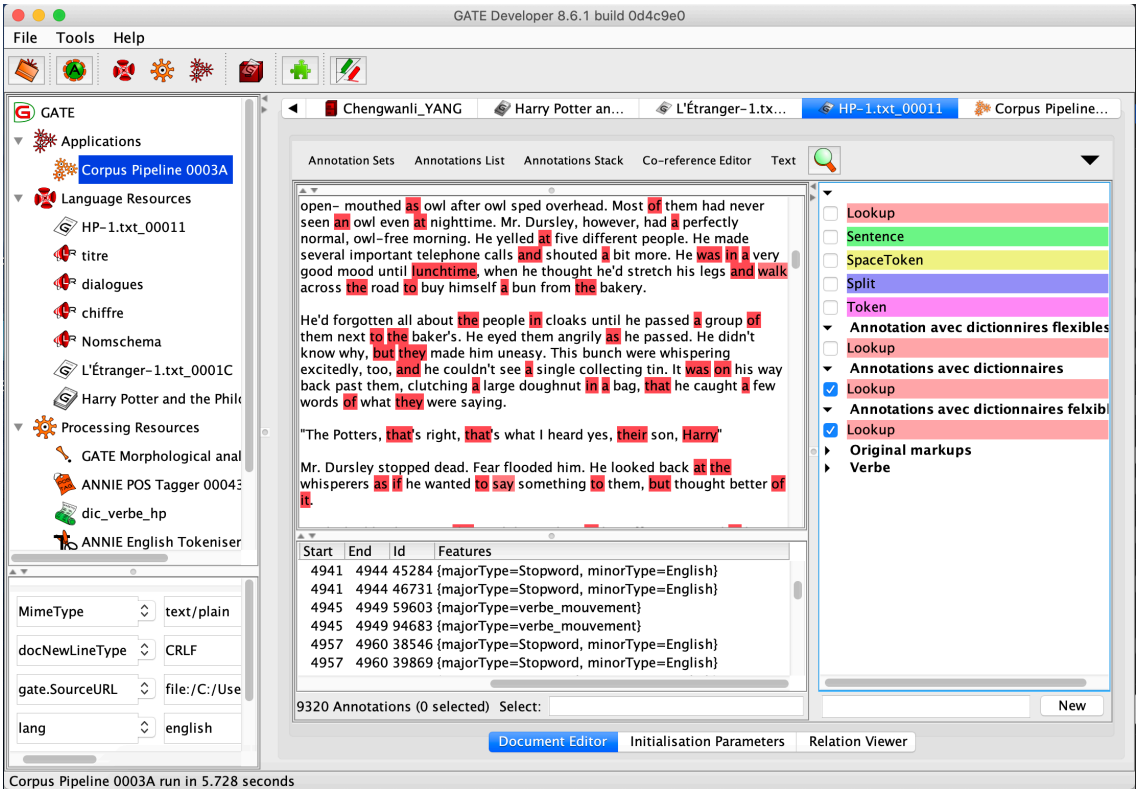


4. Annotations avec dictionnaires :

Pour le corpus « Étranger », je crée 4 dictionnaires avec Hash Gazetteer pour annoter des lieux, des meubles, des personnages (masculin et féminin) et le stopword en français commencé par l'alphabet a b c.



Pour le corpus « Harry Potter and the Philosopher's Stone », avec la même méthode, je fais la création des 4 dictionnaires « lieux » (maison, public, monde), « Personnage », « Reference_temporelle » et « Stopword ». Ensuite je crée des dictionnaires flexibles pour détecter des verbes pertinents indépendamment du temps verbal utilisé ,par exemple: verbes de mouvement (go, walk, run…) et verbes du discours (say, tell, talk…). Pour cela marche, j'utilise ANNIE English Tokeniser, ANNIE POST Tagger, GATE Morphological analyser et les dictionnaires flexibles que j'ai créés.



5. Création des grammaires JAPE :

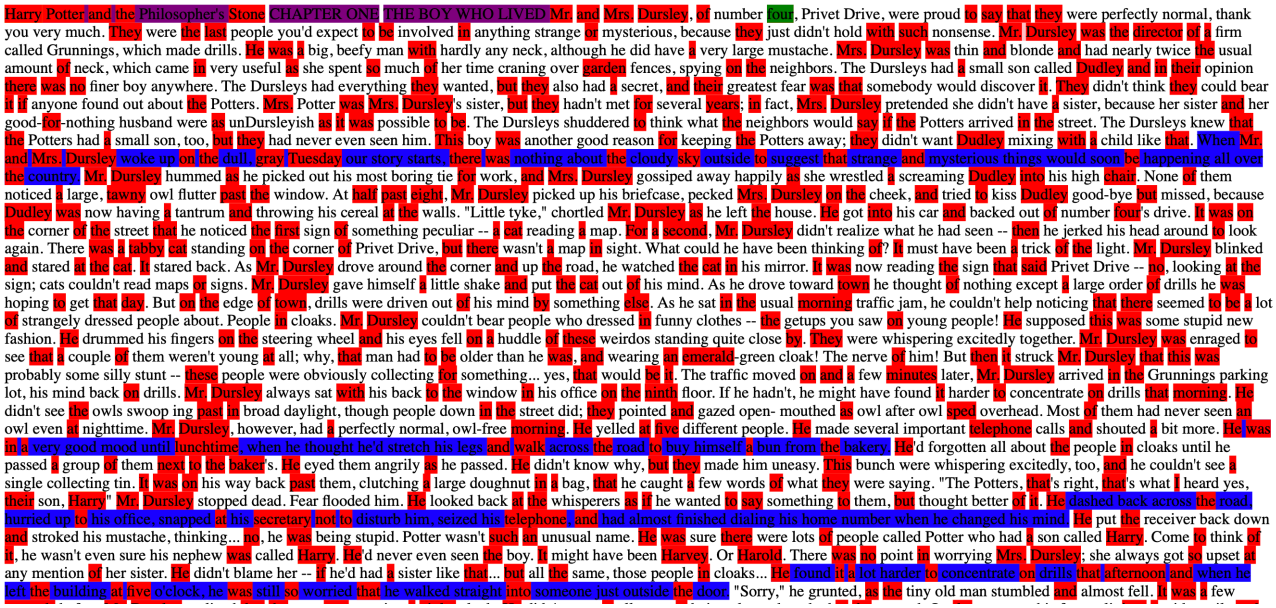
Pour le corpus « Harry Potter and the Philosopher's Stone », je crée un fichier « PhraseWHEN.jape » pour annoter des passages importants en donnant l'étiquette « PhraseWhen ». Ils contiennent le mot « when ». C'est une conjonction qui lie 2 évènements, surtout dans la description. Pour annoter le personnage et le verbe, je crée « Token-perso-verbe.jape » et « Token-verbe2.jape ». L'idée est de trouver un token (suivi le mot when) étant le personnage, et puis trouver un token ayant la catégorie VB ou VBD, c'est-à-dire le verbe, entre le personnage et le verbe, il est possible d'exister un ou plusieurs token.
Mais il y a un problème, ce traitement fonctionne partiellement. Dans plusieurs « PharseWhen » que j'ai annoté, « WhenPersonnage » et « Whenverbe » ne sont pas affichés en même temps. Autrement dit, dans la phrase, la grammaire de JAPE trouve soit « WhenPersonnage », soit « Whenverbe ».
La grammaire est suivante:

```
Phase: Phrases
Input: Token
Options: control = appelt

Rule:TokenpersoRule
(
 {Token.string==~"[Ww]hen"}
 ({Token.kind == word}):t
)
-->
:t.WhenPersonnage = {rule = "TokenpersoRule"}

Rule:TokenverbeoRule
(
 {Token.string==~"[Ww]hen"}
 {Token.kind == word}
 ({Token.category == VB} | {Token.category == VBD}):f
)
-->
:f.WhenVerbe = {rule = "TokenverbeRule"}
```

Je trouve d'abord le mot when, et le token suivante (WhenPersonnage), ensuite le verbe (VB ou VBD). Mais le résultat n'est pas suffisant.

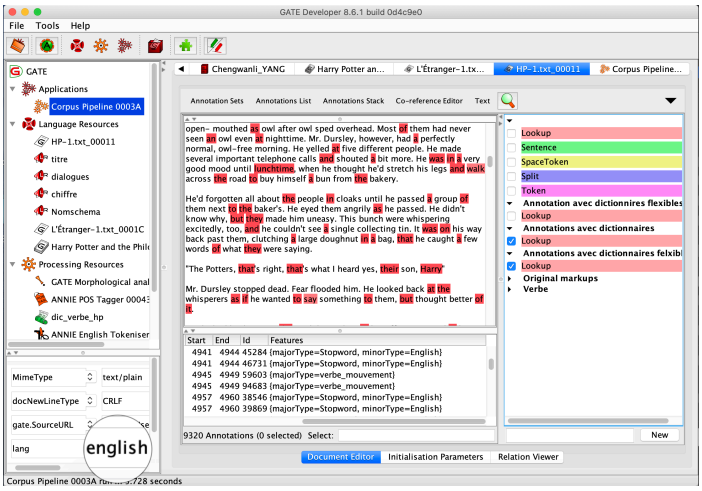
6. Impression d’annotations :

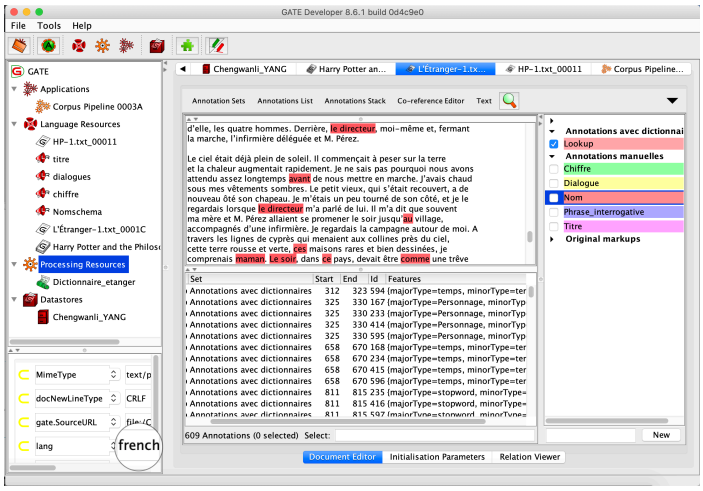
Je fais l’export toutes les annotations du 1er doucement de corpus « Harry Potter and the Philosopher's Stone » en xml, et puis copier-coller dans un fichier html. Dans le fichier css, je définis la couleur de chaque type d’annotation. Enfin je l'ouvre sur le navigateur et l’imprime en PDF.



7. Détection des langues :

Afin de détecter la langue, j’utilise TextCat Language Identification.





8. Création d’une application avec des traitements conditionnels :

A la fin, je sauvegarde l’application qui contient tous les traitement que j’ai utilisé, nommé « Application-HP.gapp ».

Conclusion

Grosso modo ce projet présente 2 façons de l'annotation: manuellement et automatiquement. Annoter manuellement permet à l'annotateur de trouver le token ou le passage plus précisément, c'est-à-dire que l'annotateur décide d'annoter quelle paragraphe ou quel token dont il a besoin. Dans le corpus « Étranger », j'ai annoté les dialogues importants au lieu d'annoter automatiquement tous les dialogues. Évidemment ce moyen n'est pas suffisant pour le grand corpus. Si toutes les annotations sont soulignées par la main, nous perdons beaucoup de temps.

En revanche, l'automaticité rend le travail efficace. Avec Hash Gazetteer nous pouvons définir le dictionnaire pour les annotations spécifiques. Par exemple dans le corpus « Harry Potter and the Philosopher's Stone », j'ai crée un dictionnaire pour trouver des références temporelles et des verbes qui ont le temps et le mode. Grâce à la grammaire JAPE, nous pouvons aussi annoter des passages importants dans le document en identifiant le personnage, le contexte et le temps. Cette méthode pratique et efface pour traiter des gros corpus, il n'y a guère d'erreurs. Mais il faut bien écrire des règles de grammaire ou dictionnaire. Pendant le travail, nous devrions utiliser ces deux façons afin d'augmenter l'efficacité et l'améliorer.