



# **Projet final**

Présenté par

**Chengwanli YANG**

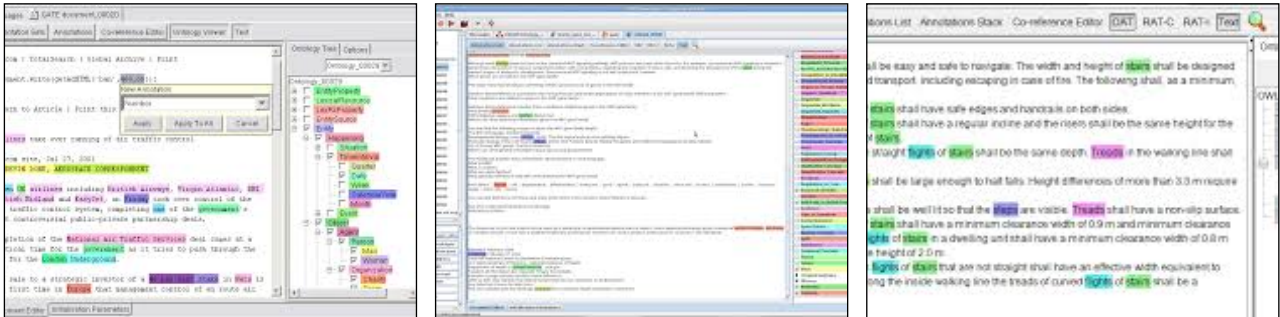
Master 1

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.



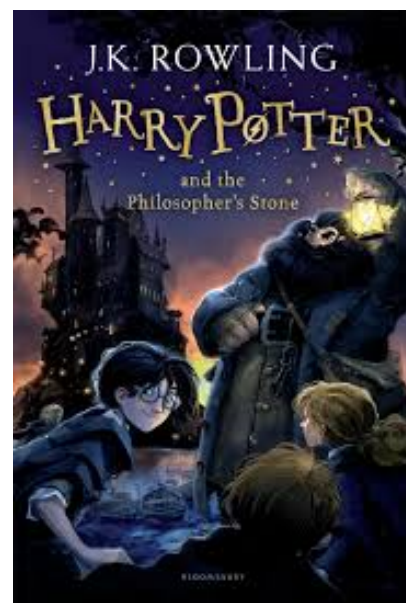
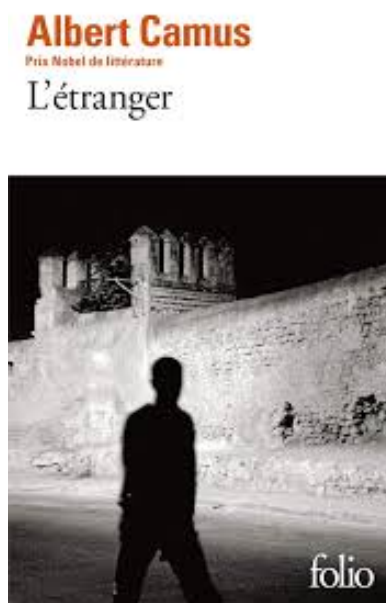
## Objectif

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.

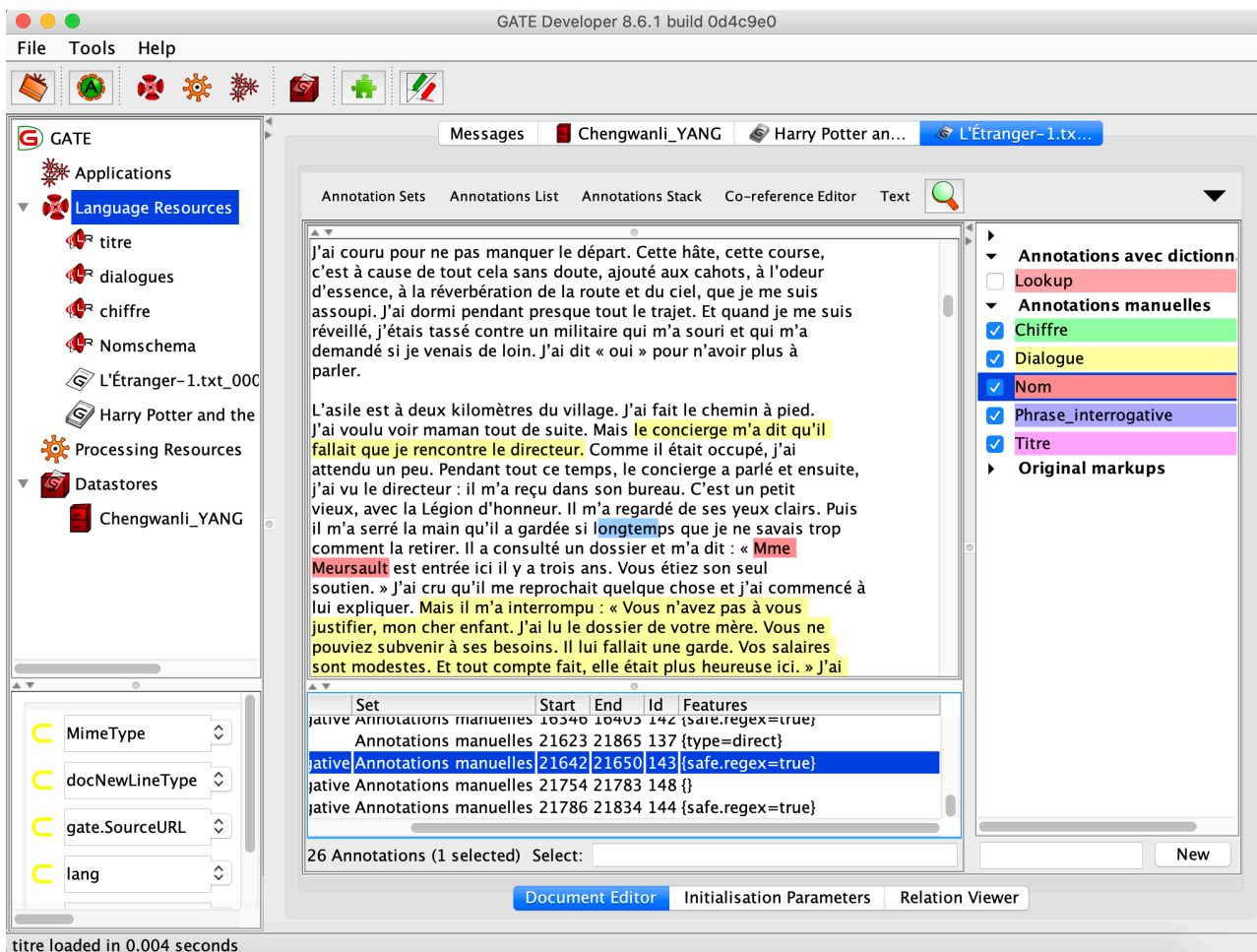


## 2. Création d'un data store :

La data store est nommé « Chengwanli\_YANG ».

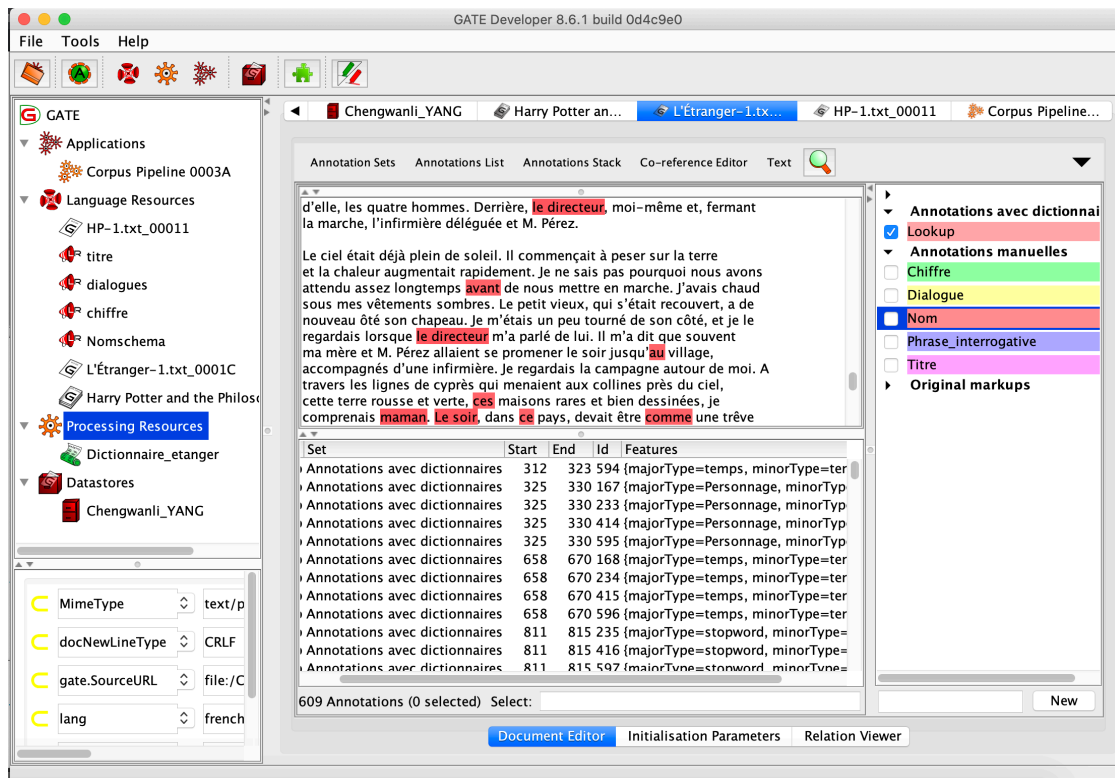
## 3. Annotations manuelles :

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'auteur, l'éditeur, le chapitre et l'année d'édition etc. De plus, je fais les annotations nommées « Phrase\_interrogative » avec l'expression régulière.

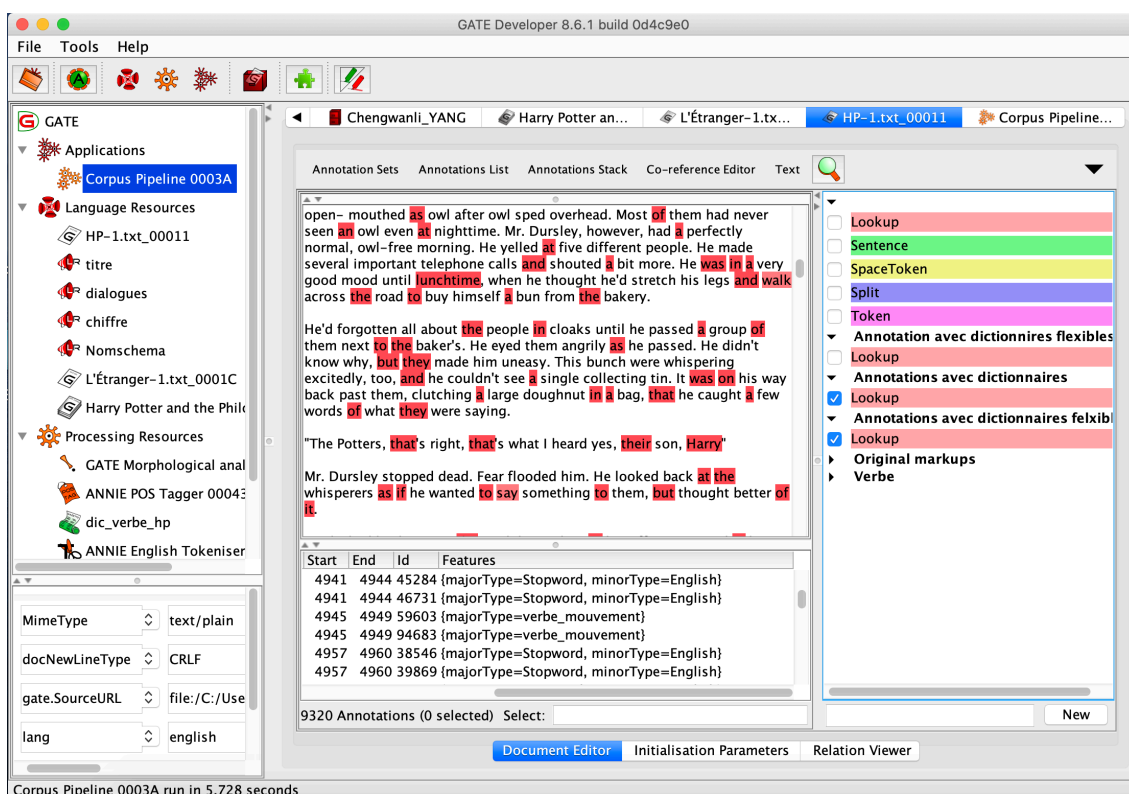


## 4. Annotations avec dictionnaires :

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 événements, 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 « PhraseWhen » 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.



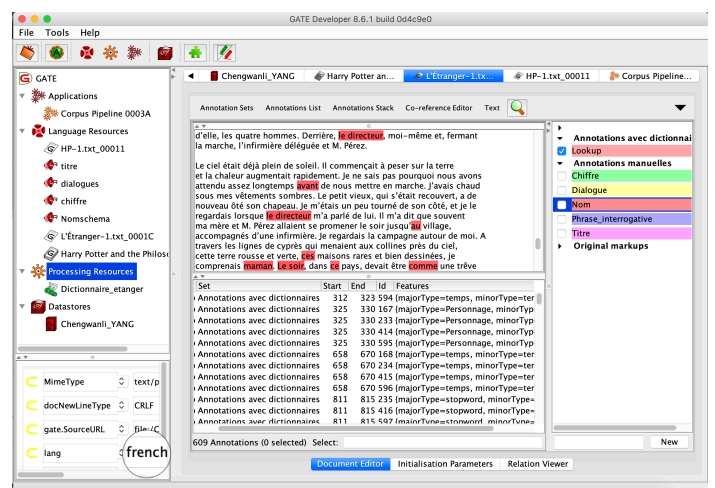
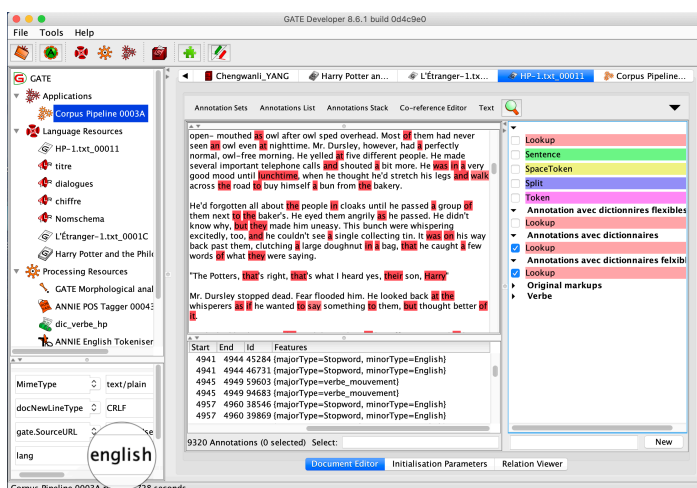
## 6. Impression d'annotations :

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.

Harry Potter and the Philosopher's Stone CHAPTER ONE THE BOY WHO LIVED Mr. and Mrs. Dursley, of number four Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as undesirable as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that. When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair. None of them noticed a large, brown owl flutter past the window. At half past eight, Mr. Dursley picked up his briefcase, pecked Mrs. Dursley on the cheek, and tried to kiss Dudley good-bye but missed, because Dudley was now having a tantrum and throwing his cereal at the walls. "Little tyke," chorled Mr. Dursley as he left the house. He got into his car and backed out of number four's drive. It was on the corner of the street that he noticed the first sign of something peculiar -- a cat reading a map. For a second, Mr. Dursley didn't realize what he had seen -- then he jerked his head around to look again. There was a tabby cat standing on the corner of Privet Drive, but there wasn't a map in sight. What could he have been thinking of? It must have been a trick of the light. Mr. Dursley blinked and stared at the cat. It stared back. As Mr. Dursley drove around the corner and up the road, he watched the cat in his mirror. It was now reading the sign that said Privet Drive -- no, looking at the sign; cats couldn't read maps or signs. Mr. Dursley gave himself a little shake and put the cat out of his mind. As he drove toward town he thought of nothing except a large order of drills he was hoping to get that day. But on the edge of town, drills were driven out of his mind by something else. As he sat in the usual morning traffic jam, he couldn't help noticing that there seemed to be a lot of strangely dressed people about. People in cloaks. Mr. Dursley couldn't bear people who dressed in funny clothes -- the getups you saw on young people! He supposed this was some stupid new fashion. He drummed his fingers on the steering wheel and his eyes fell on a huddle of these weirdos standing quite close by. They were whispering excitedly together. Mr. Dursley was enraged to see that a couple of them weren't young at all; why, that man had to be older than he was, and wearing an emerald-green cloak! The nerve of him! But then it struck Mr. Dursley that this was probably some silly stunt -- these people were obviously collecting for something... yes, that would be it. The traffic moved on and a few minutes later, Mr. Dursley arrived in the Grunnings parking lot, his mind back on drills. Mr. Dursley always sat with his back to the window in his office on the ninth floor. If he hadn't, he might have found it harder to concentrate on drills that morning. He didn't see the owls swoop in past his broad daylight, though people down in the street did; they pointed and gazed open-mouthed as owl after owl sped overhead. Most of them had never seen an owl even at nighttime. Mr. Dursley, however, had a perfectly normal, owl-free morning. He yelled at five different people. He made several important telephone calls and shouted a bit more. He was in a very good mood until lunchtime, when he thought he'd stretch his legs and walk across the road to buy himself a bun from the bakery. He'd forgotten all about the people in cloaks until he passed a group of them next to the baker's. He eyed them angrily as he passed. He didn't know why, but they made him uneasy. This bunch were whispering excitedly, too, and he couldn't see a single collecting tin. It was on his way back past them, clutching a large doughnut in a bag, that he caught a few words of what they were saying. "The Potters, that's right, that's what I heard yes, their son, Harry." Mr. Dursley stopped dead. Fear flooded him. He looked back at the whisperers. If he wanted to say something to them, but thought better of it. He dashed back across the road, hurried up to his office, snapped at his secretary not to disturb him, seized his telephone, and had almost finished dialing his home number when he changed his mind. He put the receiver back down and stroked his mustache, thinking... no, he was being stupid. Potter wasn't such an unusual name. He was sure there were lots of people called Potter who had a son called Harry. Come to think of it, he wasn't even sure his nephew was called Harry. He'd never even seen the boy. It might have been Harvey, or Harold. There was no point in worrying Mrs. Dursley; she always got so upset if any mention of her sister. He didn't blame her -- he'd had a sister like that... but all the same, those people in cloaks... He found it a lot harder to concentrate on drills that afternoon and when he left the building at five o'clock, he was still so worried that he walked straight into someone just outside the door. "Sorry," he grunted, as the tiny old man stumbled and almost fell. It was a few

## 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 créé 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 efficace 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.