
Rapport final
présenté par :
NGAUV Axel

# Projet de Sémantique Computationnelle

12 décembre 2021

Master 1 de Sciences du Langage
parcours Langue et Informatique
Méthodologie de la Recherche en Informatique

# Table des matières

# Table des figures

# 1 Introduction

L'objectif de ce projet est d'annoter manuellement et automatiquement dans GATE deux livres, un en français et un en anglais. Mais que signifie donc « annoter un texte » ? Qu'est-ce qu'une annotation, et pourquoi en fait-on ?
Comme l'explique formidablement bien Iana Atanassova, « en informatique une annotation est un commentaire, une note, une explication ou tout autre remarque externe qui peut être attachée à un document ou à une partie de celui-ci. L'annotation textuelle consiste à enrichir un texte avec des informations, rattachées aux parties du texte. L'annotation a pour but d'expliciter (ou de « traduire ») certaines propriétés des éléments textuels (notamment le sens) qui sont normalement inaccessibles pour la machine. Elle permet donc à la machine d'accéder au sens par le biais des annotations (les étiquettes attribués aux éléments textuels) qui reflètent, si elles sont correctes, une partie du sens du texte. »[Atanassova, 2011]

Pour mener à bien ce projet, j'ai donc eu recours au logiciel GATE [Cunningham et al., 2011], acronyme pour « General Architecture for Text Engineering » et à son système d'extraction d'information (GATE a été à l'origine développé pour de l'extraction d'informations) : ANNIE, pour « A Nearly-New Information Extraction System » [Cunningham et al., 2002].

Ce logiciel dispose de trois types de ressources :

1. Language Ressources (documents et corpus)
2. Processing Ressources (traitements pour l'analyse des documents)
3. Applications (des chaînes de traitements)

Ce rapport présentera dans un premier temps les différentes étapes de ce projet, puis mettra dans un second temps en lumière les forces et faiblesses des annotations manuelles et automatiques.

# 2 Description des différentes étapes

## 2.1 Sélection des livres

J'ai donc choisi deux livres : l'un en français, l'autre en anglais.
Livre en français :

— Titre : Cent ans après ou l'An 2000
— Auteur : Edward Bellamy

Livre en anglais :

— Titre : The Legends Of King Arthur And His Knights
— Auteur : James Knowles, Thomas Malory

## 2.2 Création des corpus

J'ai seulement sélectionné des extraits pour chacun des deux livres afin de constituer mes corpus, je ne les ai pas pris en entier. Corpus en français :

— Nombre de chapitres : 12
— Nombre de mots : 26 769

Corpus en anglais :

— Nombre de chapitres : 7
— Nombre de mots : 31 387

## 2.3 Création d'un data store

Ensuite, j'ai créé un data store afin d'y sauvegarder mon corpus avec les documents annotés.

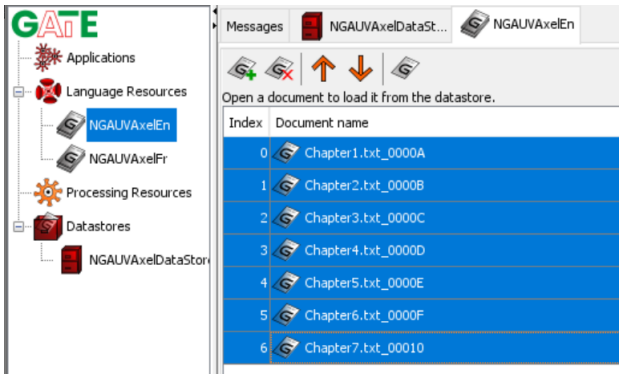


FIGURE 1 – Etape 1, 2 et 3 : sélection des livres, création des corpus et d'un data store

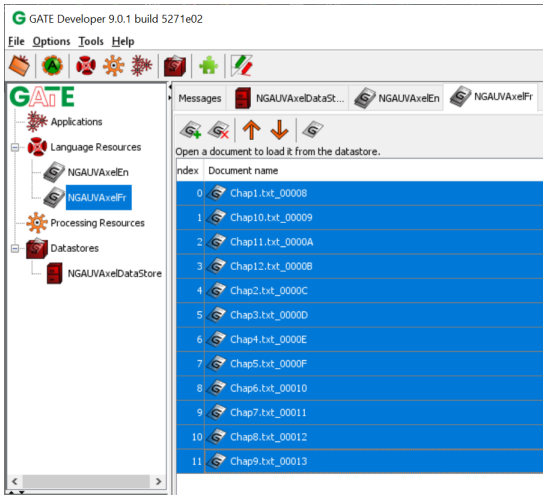


FIGURE 2 – Etape 1, 2 et 3 : sélection des livres, création des corpus et d'un data store

## 2.4 Création des schémas d'annotation

Cette étape consiste en la création de quatre schémas d'annotation. Le premier concerne les titres, le second les dialogues, le troisième les dates et le dernier les questions.

Pour le schéma d'annotation des dates, j'ai choisi de distinguer les dates selon leur format :

- Année
- Mois
- Jour (dans le mois)
- Jour de la semaine
- L'année, le mois et le jour
- L'année et le mois
- Le mois et le jour

J'ai choisi d'annoter les dates car dans cette œuvre le rapport au temps est important. En effet, le protagoniste de l'histoire effectue un voyage temporel.

Pour le schéma d'annotation des questions, j'ai fait le choix de distinguer les questions que le locuteur se pose à lui-même des questions qu'il pose à son interlocuteur. Le choix d'annoter les questions est également pertinent ici puisque le protagoniste, après son voyage dans le temps, se pose (et pose)

énormément de question. Ce qui bien sûr, est tout à fait normal.

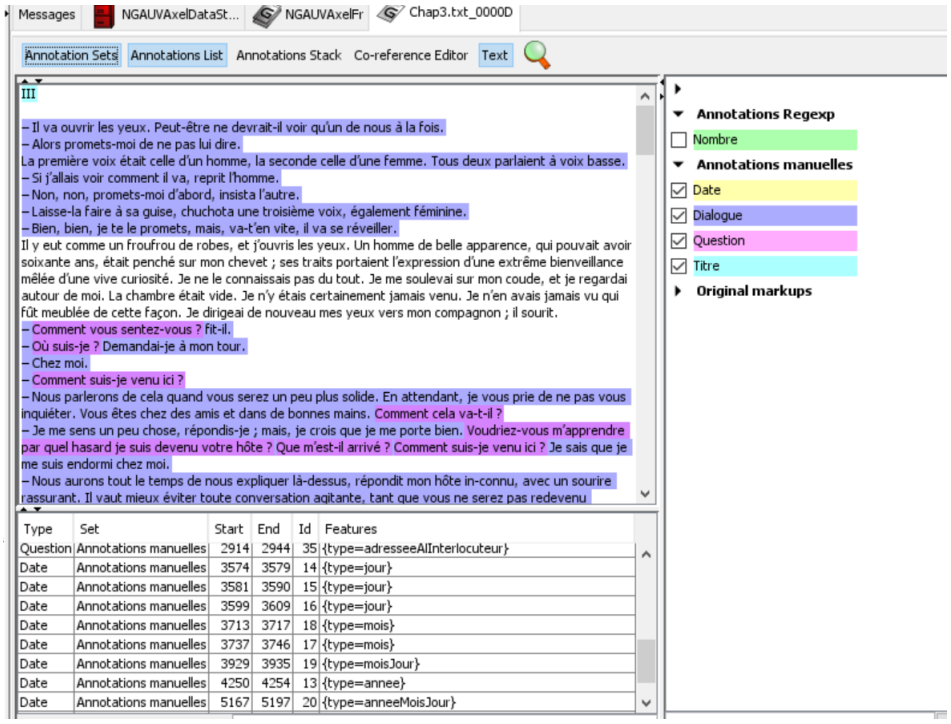


FIGURE 3 – Etape 4 : création des schémas d'annotation

Cette étape de création des schémas d'annotation a été effectué sur le chapitre 3 du corpus français. Les schémas d'annotations sont regroupés dans l'ensemble « Annotations manuelles ».

## 2.5 Création d'annotations avec expressions régulières

Il s'agit maintenant de faire des annotations à l'aide d'expressions régulières. En utilisant les expressions régulières, j'ai choisi d'annoter :

— Les nombres (toujours cette importance des dates)
— Les différents éléments de ponctuation
— Les pronoms personnels à la première personne du singulier (`\b(J'|j'Je|je|Me|me)\b`)

Cette étape de création d'annotations avec des expressions régulières a été effectué sur le chapitre 1 du corpus français. Les annotations avec expressions

régulières sont regroupées dans l'ensemble « Annotations manuelles ».

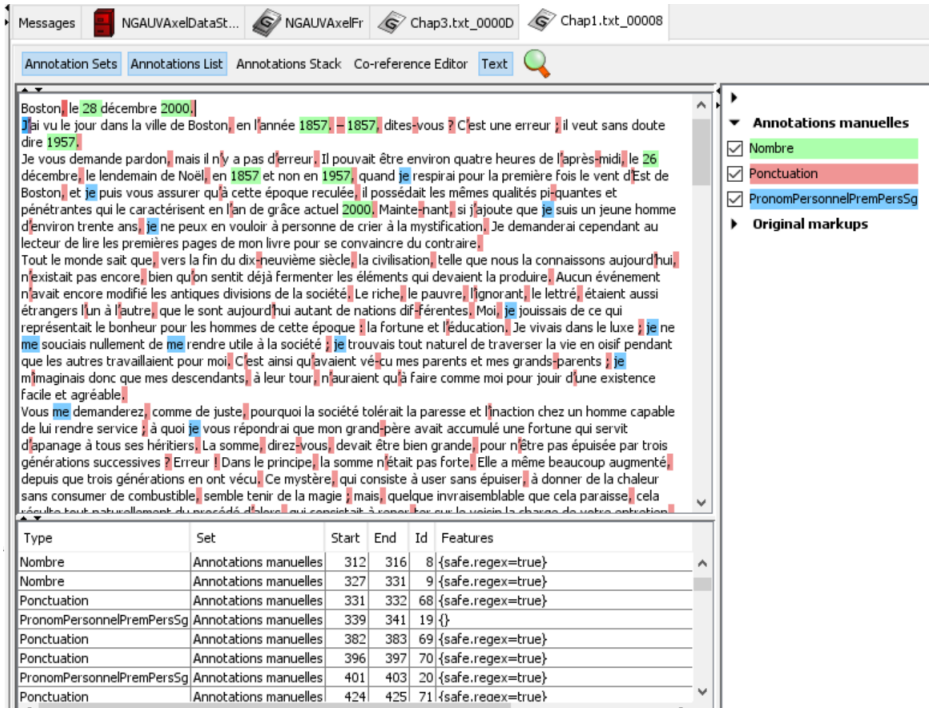


FIGURE 4 – Etape 5 : création d'annotations avec expressions régulières

## 2.6 Création de dictionnaires

J'ai créé quatres dictionnaires afin d'annoter mon corpus. A partir de cette étape, j'ai choisi de travailler avec le corpus en anglais. J'ai choisi d'annoter :

- Les personnages
- Les évènements
- Les lieux réels
- Les lieux fictifs

Pour cela, j'ai créé quatres fichiers avec l'extension « lst » contenant chacun une liste (personnages, événements, lieux réels, lieux fictifs) ainsi qu'un fichier « lists » avec l'extension « def » pour définir un ensemble de dictionnaire et leur donner des attributs.
J'ai ensuite créé un « Hash Gazetteer » à partir du fichier « lists.def » puis j'ai créé une application à partir de ce « Hash Gazetteer » que j'ai appliqué

sur mon corpus anglais. Les annotations produites sont des lookup. Je les ai regroupé dans le groupe « Annotations avec dictionnaires ».

L'œuvre qui constitue le corpus en anglais étant un récit fictif reposant sur une légende, celle du Roi Arthur, j'ai trouvé pertinent de me focaliser sur les personnages de l'histoire, les évènements qui se produisent dans cette histoire, ainsi que sur les lieux réels mais aussi fictifs.

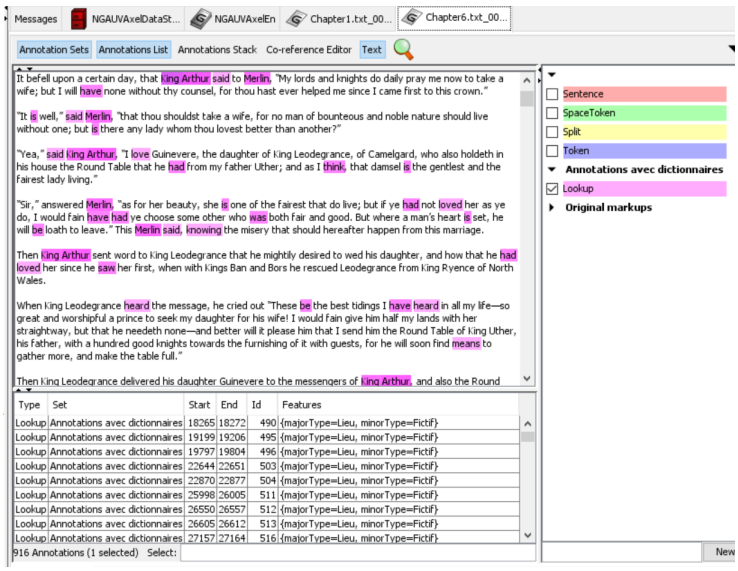


FIGURE 5 – Etape 6 : création de dictionnaires

## 2.7 Création de dictionnaires flexibles

Cette étape consiste en la création de dictionnaires flexibles. Pour cela, j'ai tout d'abord créé une analyse morphologique avec « GATE Morphological Analyser », puis j'ai créé deux listes de verbes à l'infinitif en anglais : une liste pour les verbes d'état dans le fichier « VerbesEtat.lst » et une liste pour les verbes d'action dans « VerbesAction.lst ».
Enfin, j'ai créé un fichier « listsV.def » pour définir un ensemble regroupant ces deux dictionnaires et leur donner des attributs (majorType = Verbe ; minorType = Etat ou Action).
A partir de ce fichier j'ai créé un « Hash Gazetteer ». Tout cela va me servir à créer mon « Flexible Gazetteer ». Ce dernier me permet d'annoter des verbes indépendamment du temps verbal utilisé.

Dans ce but, j'éxecute d'abord ANNIE, puis je crée une application avec un « Tokeniser », un « Sentence splitter », un « POS Tagger », un « Morphological analyser » et mon « Flexible Gazetteer ». Pour finir, j'éxecute cette application sur mon corpus en anglais.

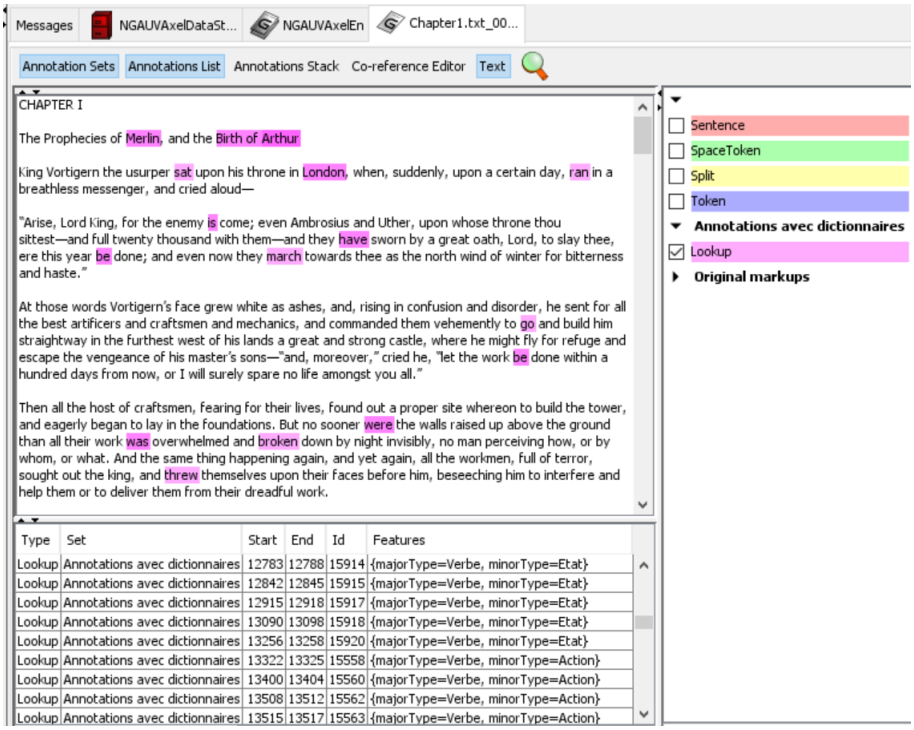


FIGURE 6 – Etape 7 : création de dictionnaires flexibles

## 2.8 Impression d'annotations

J'ai effectué l'impression d'annotations sur le chapitre 11 du corpus en français (Cent ans après ou l'An 2000). Sont annotés les dates (en vert), les dialogues (en rose), les questions (en bleu) ainsi que le titre (en marron). Ces annotations ne sont pas exhaustives : je les ai faites manuellement simplement afin d'illustrer l'impression d'annotations.

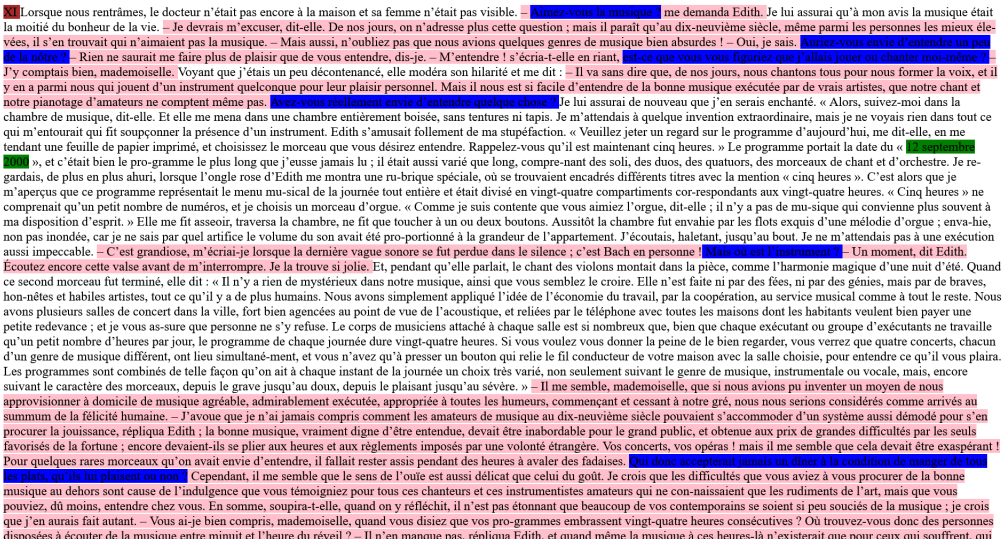
XILorsque nous rentrâmes, le docteur n'était pas encore à la maison et sa femme n'était pas visible. – Aimez-vous la musique ? me demanda Edith. Je lui assurai qu'à mon avis la musique était la moitié du bonheur de la vie. – Je devrais m'excuser, dit-elle. De nos jours, on n'adresse plus cette question ; mais il paraît qu'au dix-neuvième siècle, même parmi les personnes les mieux élevées, il s'en trouvait qui n'aimaient pas la musique. – Mais aussi, n'oubliez pas que nous avions quelques genres de musique bien absurdes ! – Oui, je sais. Auriez-vous envie d'entendre un peu de la nôtre ? – Rien ne saurait me faire plus de plaisir que de vous entendre, dis-je. – M'entendre ! s'écria-t-elle en riant, est-ce que vous vous figurez que j'allais jouer ou chanter moi-même ? – J'y comptais bien, mademoiselle. Voyant que j'étais un peu décontenancé, elle modéra son hilarité et me dit : – Il va sans dire que, de nos jours, nous chantons tous pour nous former la voix, et il y en a parmi nous qui jouent d'un instrument quelconque pour leur plaisir personnel. Mais il nous est si facile d'entendre de la bonne musique exécutée par de vrais artistes, que notre chant et notre pianotage d'amateurs ne comptent même pas. Avez-vous réellement envie de m'entendre quelque chose ? – Je lui assurai de nouveau que j'en serais enchanté. « Alors, suivez-moi dans la chambre de musique, dit-elle. Et elle me mena dans une chambre entièrement boisée, sans tentures ni tapis. Je m'attendais à quelque invention extraordinaire, mais je ne voyais rien dans tout ce qui m'entourait qui fît soupçonner la présence d'un instrument. Edith s'amusait follement de ma stupéfaction. « Veuillez jeter un regard sur le programme d'aujourd'hui, me dit-elle, en me tendant une feuille de papier imprimé, et choisissez le morceau que vous désirez entendre. Rappelez-vous qu'il est maintenant cinq heures. » Le programme portait la date du « 12 septembre 2000 », et c'était bien le pro-gramme le plus long que j'eusse jamais lu ; il était aussi varié que long, compre-nant des soli, des duos, des quatuors, des morceaux de chant et d'orchestre. Je regardais, de plus en plus ahuri, lorsque l'ongle rose d'Edith me montra une ru-brique spéciale, où se trouvaient encadrés différents titres avec la mention « cinq heures ». C'est alors que je m'aperçus que ce programme représentait le menu mu-sical de la journée tout entière et était divisé en vingt-quatre compartiments cor-respondants aux vingt-quatre heures. « Cinq heures » ne comprenait qu'un petit nombre de numéros, et je choisis un morceau d'orgue. « Comme je suis contente que vous aimiez l'orgue, dit-elle ; il n'y a pas de mu-sique qui convienne plus souvent à ma disposition d'esprit. » Elle me fit asseoir, traversa la chambre, ne fit que toucher à un ou deux boutons. Aussitôt la chambre fut envahie par les flots exquis d'une mélodie d'orgue ; enva-hie, non pas inondée, car je ne sais par quel artifice le volume du son avait été pro-portionné à la grandeur de l'appartement. J'écoutais, haletant, jusqu'au bout. Je ne m'attendais pas à une exécution aussi impeccable. – C'est grandiose, m'écriai-je lorsque la dernière vague sonore se fut perdue dans le silence ; c'est Bach en personne ! Mais où est l'instrument ? – Un moment, dit Edith. Écoutez encore cette valse avant de m'interrompre. Je la trouve si jolie. Et, pendant qu'elle parlait, le chant des violons montait dans la pièce, comme l'harmonie magique d'une nuit d'été. Quand ce second morceau fut terminé, elle dit : « Il n'y a rien de mystérieux dans notre musique, ainsi que vous semblez le croire. Elle n'est faite ni par des fées, ni par des génies, mais par de braves, hon-nêtes et habiles artistes, tout ce qu'il y a de plus humains. Nous avons simplement appliqué l'idée de l'économie du travail, par la coopération, au service musical comme à tout le reste. Nous avons plusieurs salles de concert dans la ville, fort bien agencées au point de vue de l'acoustique, et reliées par le téléphone avec toutes les maisons dont les habitants veulent bien payer une petite redevance ; et je vous as-sure que personne ne s'y refuse. Le corps de musiciens attaché à chaque salle est si nombreux que, bien que chaque exécutant ou groupe d'exécutants ne travaille qu'un petit nombre d'heures par jour, le programme de chaque journée dure vingt-quatre heures. Si vous voulez vous donner la peine de le bien regarder, vous verrez que quatre concerts, chacun d'un genre de musique différent, ont lieu simultané-ment, et vous n'avez qu'à presser un bouton qui relie le fil conducteur de votre maison avec la salle choisie, pour entendre ce qu'il vous plaira. Les programmes sont combinés de telle façon qu'on ait à chaque instant de la journée un choix très varié, non seulement suivant le genre de musique, instrumentale ou vocale, mais, encore suivant le caractère des morceaux, depuis le grave jusqu'au doux, depuis le plaisant jusqu'au sévère. » – Il me semble, mademoiselle, que si nous avions pu inventer un moyen de nous approvisionner à domicile de musique agréable, admirablement exécutée, appropriée à toutes les humeurs, commençant et cessant à notre gré, nous nous serions considérés comme arrivés au summum de la félicité humaine. – J'avoue que je n'ai jamais compris comment les amateurs de musique au dix-neuvième siècle pouvaient s'accommoder d'un système aussi démodé pour s'en procurer la jouissance, répliqua Edith ; la bonne musique, vraiment digne d'être entendue, devait être inabordable pour le grand public, et obtenue aux prix de grandes difficultés par les seuls favorisés de la fortune ; encore devaient-ils se plier aux heures et aux règlements imposés par une volonté étrangère. Vos concerts, vos opéras ! mais il me semble que cela devait être exaspérant ! Pour quelques rares morceaux qu'on avait envie d'entendre, il fallait rester assis pendant des heures à avaler des fadaises. Qui donc accepterait jamais un dîner à la condition de manger de tous les plats, qu'ils lui plaisent ou non ? Cependant, il me semble que le sens de l'ouïe est aussi délicat que celui du goût. Je crois que les difficultés que vous aviez à vous procurer de la bonne musique au dehors sont cause de l'indulgence que vous témoigniez pour tous ces chanteurs et ces instrumentistes amateurs qui ne con-naissaient que les rudiments de l'art, mais que vous pouviez, dû moins, entendre chez vous. En somme, soupira-t-elle, quand on y réfléchit, il n'est pas étonnant que beaucoup de vos contemporains se soient si peu souciés de la musique ; je crois que j'en aurais fait autant. – Vous ai-je bien compris, mademoiselle, quand vous disiez que vos pro-grammes embrassent vingt-quatre heures consécutives ? Où trouvez-vous donc des personnes disposées à écouter de la musique entre minuit et l'heure du réveil ? – Il n'en manque pas, répliqua Edith, et quand même la musique à ces heures-là n'existerait que pour ceux qui souffrent, qui

FIGURE 7 – Etape 8 : impression d'annotations

## 2.9 Détection des langues

Il s'agit maintenant de détecter la langue des textes de nos corpus et de l'ajouter en tant que caractéristique du document. Nous avons vu en cours que pour atteindre cet objectif nous pouvions utiliser le composant « LingPipe ». Cependant, je ne dispose pas de « LingPipe ». J'ai cherché dans la liste de « CREOLE Plugin Manager » et même utilisé sa barre de recherche, sans succès. Je me suis donc rabattu sur le composant « Language Identification » afin d'utiliser « TextCat Language Identification », ce qui m'a permis d'atteindre l'objectif recherché.

Les textes composant le corpus en français ont bien été détecté comme étant en français et les textes composant le corpus en anglais comme étant en anglais. Cela a été rajouté en tant que caractéristique pour chaque document.

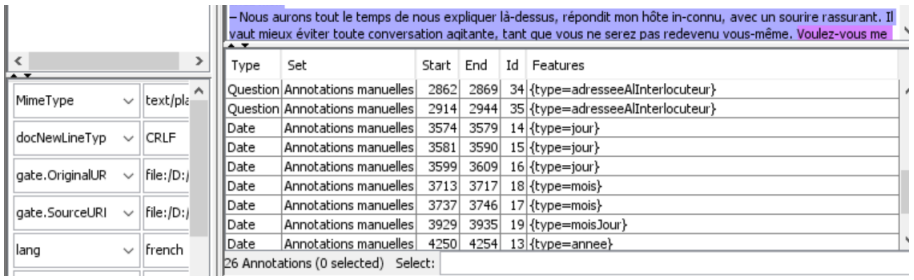


FIGURE 8 – Etape 9 : détection de la langue sur le corpus en français

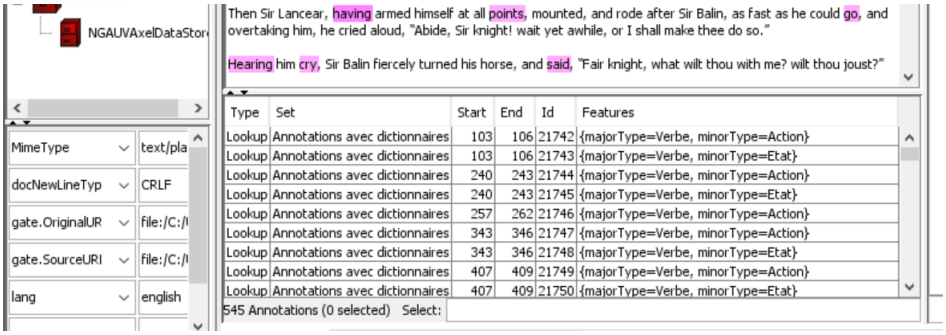


FIGURE 9 – Etape 9 : détection de la langue sur le corpus en anglais

## 2.10 Rédaction du rapport

Cette dernière étape consiste en la rédaction du rapport concernant le projet. Le rapport en question est ce présent document.
Après avoir présenté les différentes étapes du projet et ce que j'ai acccompli, je vais maintenant mettre en lumière les forces et faiblesses des annotations manuelles et automatiques.

# 3 Forces et faiblesses des annotations manuelles et automatiques

Avec les cours de Sémantique Computationnelle [Eyharabide, 2021] que j'ai suivi et le projet qu'il nous a été demandé de réaliser, j'ai pu m'apercevoir des forces et faiblesses des annotations manuelles et automatiques.

## 3.1 Les forces des annotations manuelles

Les forces des annotations manuelles sont :
- Moins d'erreurs d'annotations qu'avec les annotations automatiques lorsque le texte contient des erreurs ou des éléments de langage qui n'appartiennent pas à la langue standard
- Nul besoin d'avoir recours à des règles ou à des algorithmes, car l'annotateur (contrairement à la machine) comprend de lui-même le texte à annoter

## 3.2 Les faiblesses des annotations manuelles

Les annotations manuelles ont plusieurs faiblesses :
- Elles sont chronophages (les annotations manuelles sont faites par l'annotateur)
- Plus la tâche est grande ou à répéter, et plus le coût en temps est important
- Dans le cas des annotations avec expressions régulières, possibilité d'erreur de la part de l'annotateur lors de la conception des expressions régulières

## 3.3 Les forces des annotations automatiques

Les forces des annotations automatiques sont :
- Gain de temps car la tâche se fait automatiquement
- Plus la tâche est grande ou à répéter, et plus l'intérêt pour les annotations automatiques est grand

## 3.4 Les faiblesses des annotations automatiques

La possibilité d'erreurs dans les annotations (une annotation est dite erronée lorsque sa valeur diffère de celle attribuée par un expert). Ces erreurs peuvent être dûes à :

— Un algorithme erroné qui serait incapable (en partie ou totalement) d'attribuer les bonnes catégories aux éléments textuels
— Une absence de définition formelle pour chaque catégorie d'annotation (la tâche étant automatique, la machine doit pouvoir déterminer pour chacun des élément s'il appartient à la catégorie ou non)
— La source même (le texte peut contenir des erreurs ou des éléments de langage qui n'appartiennent pas à la langue standard, ils peuvent donc ne pas être reconnus)

Il peut y avoir également tout simplement une absence d'annotation (un élément textuel devrait être annoté mais ne l'est pas). Les causes de cette incapacité à pouvoir identifier automatiquement les éléments textuels à annoter peuvent être les mêmes que celles citées ci-dessus.

# Références

[Atanassova, 2011] Atanassova, I. (2011). Annotation sémantique et ressources linguistiques. pages 3–5.

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

[Cunningham et al., 2011] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

[Eyharabide, 2021] Eyharabide, M. V. (2021). Sémantique computationnelle.