

Rapport projet Gate

Introduction

L'objectif de ce projet était d'explorer la puissance des différents outils tout en appliquant ce que nous avons appris en cours afin de pouvoir exploiter des corpus de texte.

Pour ce projet, j'ai choisi deux livres. Le premier, *Germinal*, écrit par Emile Zola et classique de la littérature française. Le second, *Of Mice and Men*, roman de John Steinbeck qui, par son style, pourrait presque apparaître comme une pièce de théâtre étant donné que la majorité du texte consiste en dialogues entre les deux protagonistes.

Afin d'exploiter ces données, j'ai d'abord séparé le texte en chapitres afin d'avoir 5 documents distincts pour chaque œuvre.

Ensuite, j'ai créé deux corpus (Click droit : Language Resources → New Gate Corpus), dans ces corpus, j'ai intégré mes 5 chapitres à ces corpus (Click droit : Populate → URL de mon dossier chapitres).

Afin de conserver mes données et de pouvoir les réutiliser, j'ai créé un dossier « ProjetDataStore ». Ensuite j'ai créé mon datastore dans Gate (SerieDataStore) en sélectionnant mon dossier précédemment créé.

Ce rapport décrira le déroulement de la réalisation de mon projet ainsi qu'une observation des résultats.

Table des matières

Introduction	1
I. Annotations manuelles.....	3
1. Germinal.....	3
a) Simple.....	3
b) Regex.....	3
c) Annotations schéma	4
2. Of Mice and Men.....	4
d) Simple.....	4
e) Regex.....	5
3. Conclusion annotations manuelles.....	5
II. Annotations automatiques.....	6
1. Germinal.....	6
a) Hash Gazetteer	6
b) Flexible Gazetteer	6
2. Of Mice and Men.....	8
a) Hash Gazetteer	8
b) Flexible Gazetteer	8
3. Résultats.....	8
III. Annotations JAPE.....	9
IV. Divers.....	10
1. Impression d’annotation	10
2. Détection de langue et Création d’une application avec des traitements conditionnels	11

I. Annotations manuelles

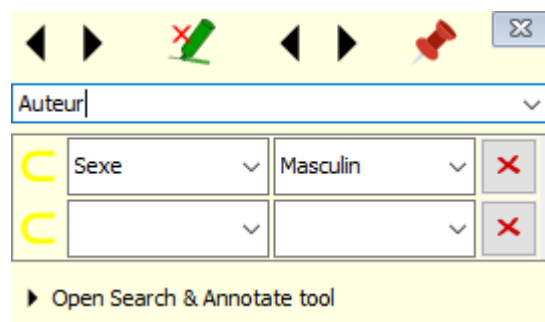
NOTE : J'ai perdu l'intégralité de mes annotations manuelles en utilisant par accident le Document Reset, par conséquent, je les ai refaites sur le chapitre 1 de chaque ouvrage.

1. Germinal

a) Simple

Dans le chapitre 1 de Germinal, j'ai annoté manuellement les Dialogues, les lieux, les numéro de chapitre, titre et auteur.

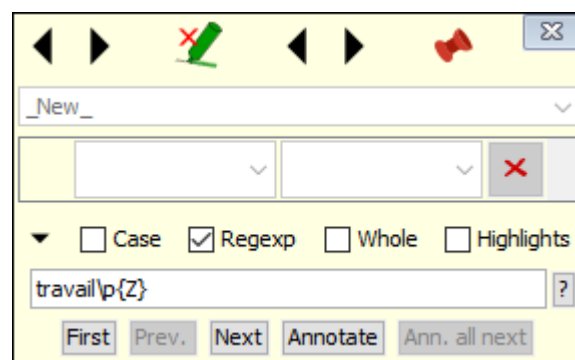
Afin d'annoter manuellement et simplement, il suffit de sélectionner la partie à annoter, click droit, choisir le nom de l'annotation. On peut également ajouter des propriétés à ces annotations

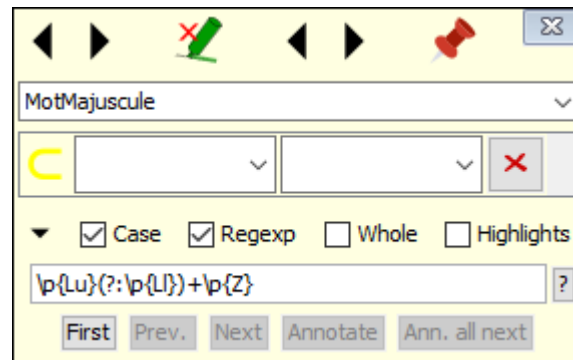


b) Regex

Pour les Regex, j'ai choisi deux exemples afin de montrer les possibilités de cet outil. Tout d'abord, j'ai choisi d'annoter le mot « travail », ensuite j'ai choisi d'annoter tous les mots commençant par une majuscule.

Pour ce faire, on utilise la même fenêtre que pour les annotations simples, on ouvre le menu contextuel « Open Search & Annotate tool » en cochant la case Regexp. Ensuite il faut cliquer sur First → Annotate → Annotate all next



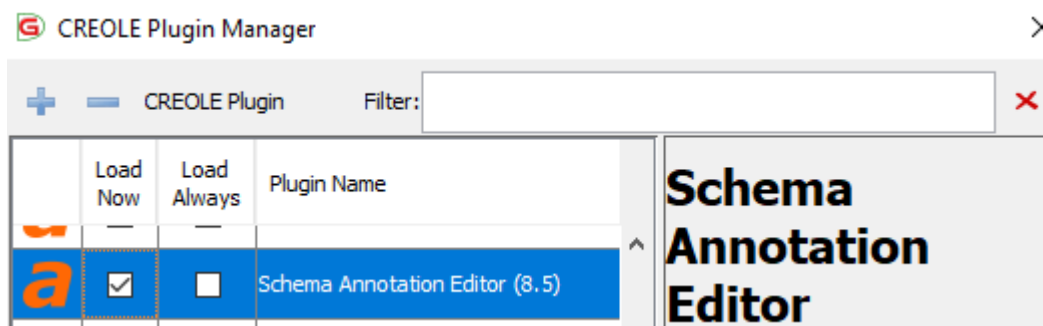


Pour la regex du mot en majuscules il faut faire bien attention de cocher la case « Case » de manière que la recherche prenne en compte la casse. Si elle n'est pas cochée, cette expression trouve l'intégralité des mots du texte.

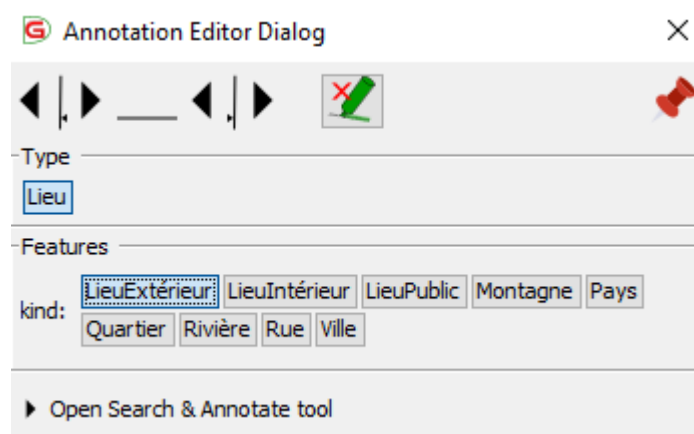
c) Annotations schéma

J'ai annoté à l'aide de schéma des lieux dans Germinal. Voir chapitre 1.

Afin d'annoter à l'aide de schéma XML, il suffit de les ajouter dans les language ressources et de les exécuter. Il faut prêter attention à bien cocher la case Schema Annotation editor dans Créole.



Les schéma XML permettent d'obtenir un « menu » avec des options afin d'annoter avec plus de précisions.



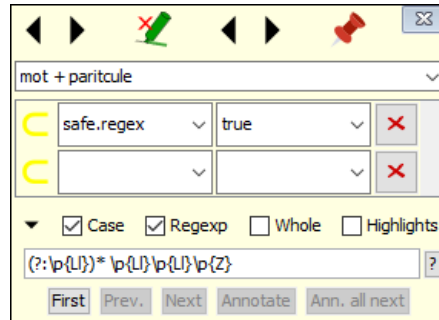
2. Of Mice and Men

d) Simple

Annotations de la même manière que pour Germinal si ce n'est que je n'ai annoté que les dialogues et le titre.

e) Regex

J'annoté en Regex toutes les occurrences du mot « mouse » dans le chapitre 1. Ensuite j'ai cherché tous les mots en minuscules suivis d'une particule. Attention, cette expression annote « Come in » sans prendre en compte le 'C' majuscule car il n'y a pas de séparateur. J'ai voulu conserver cette annotation afin de mettre en valeur le fait que les expressions régulières sont parfois à double tranchant. On pense obtenir un résultat, mais on obtient des indésirables qui les polluent. Il faut donc faire très attention et faire des tests.

The image shows a screenshot of a web-based annotation tool. At the top, there are navigation arrows and a search icon. Below that is a dropdown menu labeled 'mot + particule'. There are two rows of input fields, each with a yellow 'C' icon on the left, a dropdown menu, and a 'true' value, followed by a red 'X' icon. Below these are checkboxes for 'Case', 'Regex', 'Whole', and 'Highlights'. The 'Case' and 'Regex' checkboxes are checked. At the bottom, there is a text input field containing the regex pattern '(?:\p{L})*\p{L}\p{L}\p{Z}' and a question mark icon. Below the input field are buttons for 'First', 'Prev.', 'Next', 'Annotate', and 'Ann. all next'.

3. Conclusion annotations manuelles

Les annotations manuelles prennent beaucoup de temps sur des occurrences très mais permettent de spécifier ce qu'une annotation automatique aura du mal à trouver. Pour des occurrences faibles, elles sont idéales (par exemple le titre).

De plus, on peut facilement faire des erreurs (prendre une lettre en trop, un mot en trop, un espace). C'est assez facile lorsque la mise en page permet de repérer les éléments qu'on désire annoter, mais bien plus difficile lorsqu'on cherche des occurrences qui varient et qui sont dissimulées dans le texte par exemple, si on avait voulu annoter tous les pronoms manuellement, cela aurait pris une éternité.

Il y a donc un réel problème de régularité d'annotation : exemple est-ce que j'annote avec/sans guillemets ? Est-ce que j'annote les commentaires entre chaque citation ?

Les expressions régulières permettent de trouver rapidement une structure de mot ou d'expression sans avoir besoin d'avoir recours à une grammaire Jape ou un dictionnaire très long.

Pour conclure, l'annotation manuelle favorise grandement l'inconsistance et n'est pas adaptée pour les multiples occurrences, comme on a pu le voir lorsqu'on a annoté les nombreux dialogues de l'ouvrage de John Steinbeck. Cependant, pour des occurrences faibles, elles sont très rapides et très faciles à utiliser. Les Regex permettent manuellement de repérer des structures spécifiques, cependant, il faut être à l'aise avec cet outil et il peut arriver d'avoir des indésirables.

II. Annotations automatiques

Les annotations automatiques m'ont permis d'annoter l'intégralité des documents de mes corpus rapidement. Dans cette partie, j'expliquerai ma démarche et mes choix concernant mes annotations automatiques

1. Germinal

a) Hash Gazeteer

Les Hash Gazeteer sont simplement une liste de mots établie.

Afin d'utiliser les Hash Gazeteer, nous avons besoin de deux choses : un ou plusieurs *.lst et un *.def par dictionnaire. Les *.lst représentent mes listes de mots tandis que le *.def me permet d'associer chacun de mes *.lst à un MajorType et un MinorType.

Pour Germinal, j'ai cherché les adverbes (le MajorType) que j'ai séparé entre sous-catégories (le minorType).

Donc, j'ai ensuite fait un corpus pipeline avec mes quatre dictionnaires et l'ai exécuté sur mon corpus.

Mes quatre dictionnaires :

- Adverbes
- Personnages
- Lieux
- Indicateurs temporels

Concernant la conception de mes listes :

J'ai tout d'abord travaillé à leur élaboration.

Concernant les Adverbes : je suis allée chercher une liste des adverbes les plus utilisés en langue française et les ai classés en sous catégories.

J'ai donc obtenu 3 listes pour chaque type d'adverbe. Chacune de ces listes a été attribuée à un MinorType (Type d'adverbe) et un MajorType (Adverbe).

J'ai utilisé la même méthode pour les indicateurs temporels

Concernant les lieux, j'ai travaillé différemment sur Germinal et sur Of Mice and men. Pour Germinal, j'ai choisi de sélectionner les occurrences des lieux dans le premier chapitre afin d'observer s'ils étaient répétés dans les chapitres suivants.

Concernant les personnages, j'ai créé deux listes, une de personnages principaux, l'autre secondaire afin de détecter chacun des personnages.







Concernant les dictionnaires, j'ai associé chacune des listes à un MajorType. Les dictionnaires sont définis dans un *.def qui sera utilisé en entrée du Gate Hash Gazeteer.

J'ai également essayé d'exécuter le dictionnaire de base de Créole pour mes tests.







b) Flexible Gazeteer

Les flexible Gazeteer sont des outils très intéressants et bien plus puissants que de simples dictionnaires. Ils sont formés de la même manière que les dictionnaires, nous avons besoin d'une liste et d'une définition. Cependant leur fonctionnement est quelque peu différent, car en les combinant avec le POS Tagger, et l'English Tokenizer, nous parvenons à analyser une liste de mot et ses flexions.

Tout d'abord, on importe Annie, ensuite on crée un corpus pipeline dans lequel on intégrera le Document Reset PR (il n'est pas obligatoire et peut être embêtant car il supprime les annotations manuelles) qui va réinitialiser les annotations, un tokenizer qui permettra de d'obtenir les Tokens ainsi que les Space Token. Le sentence splitter permettant de séparer le texte en phrases. Le POS tagger qui attribuera une étiquette grammaticale à chaque mot (Attention, c'est un outil élaboré pour la langue anglaise, par conséquent, nous avons des erreurs d'étiquetage pour le texte en français). Un analyseur morphologique et enfin notre flexible Gazetteer.

Selected Processing resources		
Name	Type	
 Document Reset PR	Document Reset PR	
 ANNIE English Tokeniser	ANNIE English Tokeniser	
 ANNIE Sentence Splitter	ANNIE Sentence Splitter	
 ANNIE POS Tagger	ANNIE POS Tagger	
 Morph	GATE Morphological analyser	
 monFlexibleGazeteer	Flexible Gazetteer	

Observons notre flexible Gazetteer

Corpus:  Of Mice and Men			
Runtime Parameters for the "monFlexibleGazeteer" Flexible Gazetteer:			
Name	Type	Required	Value
 gazetteerInst	Gazetteer	✓	 VerbesOMAM
 inputASName	String		
 inputFeatureNames	List	✓	[Token.root]
 outputASName	String		AnnotationsDictionnaireFlexibleGazeteer

Ici on voit que mon dictionnaire de référence (mes listes de verbes attribuées dans ma définition) s'appelle VerbesOMAM. Mon dictionnaire flexible prend en entrée mon Token.root (une étiquette résultant de mon analyse morphologique). Il s'agit de la racine du mot qu'on peut voir ici :

Start	End	ID	Features
0	4	40397	{category=NNP, kind=word, length=4, orth=upperInitial, root=john, stri
5	14	40399	{category=NNP, kind=word, length=9, orth=upperInitial, root=steinbeck

Concernant la mise en place pour Germinal, il s'agit de la même cependant les résultats ne sont pas satisfaisants étant donné que nos Tokens et notre POS tagger sont fait pour la langue anglaise. Par conséquent, les flexions ne sont pas trouvées.

2. Of Mice and Men

a) Hash Gazetteer

Pour l'application des dictionnaires de l'ouvrage en anglais, le fonctionnement est le même, j'ai donc réalisé mes dictionnaires de verbe d'action et de discours afin d'annoter mes différents verbes sur l'ensemble de mon corpus. Les résultats sont satisfaisants cependant, nous n'avons pas les flexions de ces différents verbes. C'est là que le flexible Gazetteer entre en jeu et nous permettra d'observer des annotations bien plus précises.

b) Flexible Gazetteer

Le flexible Gazetteer m'a permis d'observer l'abondance mes verbes choisis préalablement. Ces verbes me semblaient être pertinents. J'ai pris les plus répandus afin d'augmenter la densité de mes résultats et d'observer la multiplication des occurrences en fonction du chapitre dans lequel je me trouve.

3. Résultats

En utilisant les flexible gazeteers, j'ai pu voir à la fois la limite du POS tagger (la langue) et ses forces, tout particulièrement sur mon ouvrage en anglais.

Il m'a permis d'observer que les verbes d'actions sont omniprésents dans le livre Of Mice and Men, malgré une abondance de dialogues. Je peux donc émettre l'hypothèse, sans même avoir à lire mes chapitres, que mes personnages ne sont pas statique, qu'ils se déplacent, vivent des aventures et ne sont pas uniquement dans une discussion statique ou argumentative.

J'ai pu également observer que les verbes de discours se multiplient.

Cependant, je pourrais me poser la question de la pertinence de mes listes. Peut-être sont-elles trop biaisées ? J'ai tenté d'observer les ouvrages afin de les rédiger tout en m'appuyant sur les mots les plus utilisés.

Il est intéressant de se demander si ce choix était valide étant donné que certains verbes d'actions font partie des mots les plus employés de la langue tels que go ou encore come.

III. Annotations JAPE

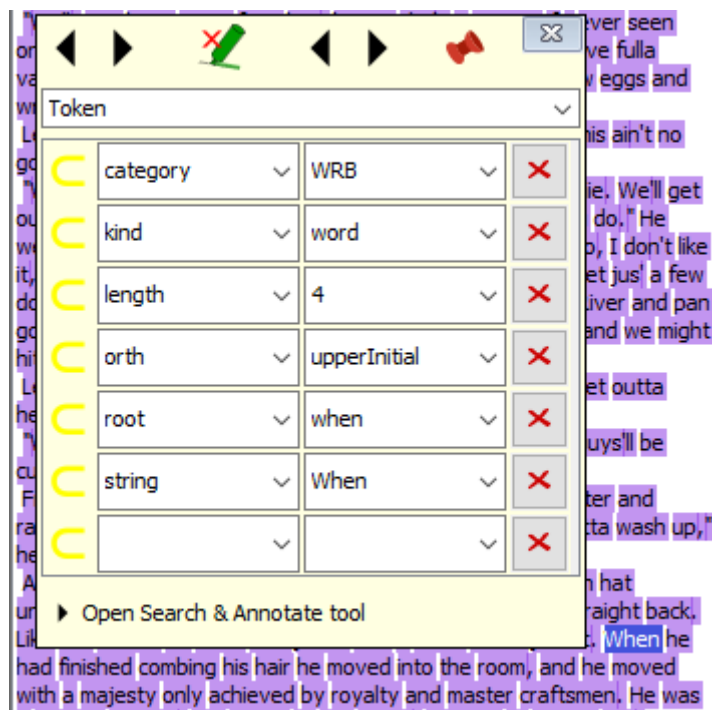
NOTE : J'ai eu beaucoup de soucis d'exécution pour mes grammaires JAPE (ordinateur qui plantait à l'exécution, particulièrement sur le corpus complet, j'ai donc fait des tests sur uniquement un documents)

Etant donné les problèmes d'analyse morphologique liée à la langue, j'ai écrit les grammaires Jape uniquement pour le document anglais.

Il aurait été possible de rechercher une phrase spécifique en utilisant {Sentence contains Token.string = « monMot »} mais je n'aurais pas annoté l'intégralité des phrases qui suivent la structure de « Quand le feu a commencé, Marie a décidé de ... ».

Afin de créer mes grammaires JAPE pour mes phrases, j'ai du tout d'abord analysé la structure de la phrase. Les phrases citées en exemple ont toutes un adverbe de temps placé soit en début de phrase, soit après la virgule. Toutes ces phrases ont une virgule. Dans chaque phrase, on retrouve le nom d'un personnage ainsi qu'un verbe d'action.

Pour commencer, j'ai créé une grammaire JAPE pour détecter chaque phrase avec cette structure. Pour ce faire, j'ai utilisé les catégories de mots attribuées par mon POS tagger



On remarque par exemple que les marqueurs de temps sont catégorisés WRB, ce qui va beaucoup nous aidé à repérer ces phrases.

Pour ma part, j'ai choisi d'annoter les phrases ayant cette structure :

« Quand., George décida de ... »

Ma phrase s'articule donc de cette manière :

[WRB] –des tokens-- , [NNP (nom propre)] [VB] –des tokens—

Dans mon dossier, il s'agit du document TrouverPhraseAction2.

De cette manière j'ai pu isoler les phrases qui m'intéressaient, cependant les résultat ne sont pas toujours à la hauteur de mes attentes. J'annote sans le vouloir des phrases telles que « Why, George started [...] » ce qui ne m'intéresse pas.

Une fois mes phrases isolées, j'ai créé des grammaires pour mes Personnages (en utilisant « NNP ») et mes verbes (en utilisant les différentes dénominations des « VB »).

Cependant, je ne suis pas parvenue exactement à ce que je voulais. J'ai essayé de retourner mon personnage et mon verbe dans deux grammaires différentes en conservant la même structure de phrase mais mes résultats varient.

IV. Divers

1. Impression d'annotation

Afin d'imprimer mes annotations, j'ai exporté mon document (chapitre 1 de Of Mice and Men) au format inline XML

```
1 John Steinbeck's
2 <Titre gate:gateId="108355">Of
3 Mice and Men </Titre>
4 CHAPTER 1
5 A FEW MILES <MotPart gate:gateId="108491">south of </MotPart>Soledad, the Salinas River <MotPart
gate:gateId="108493" safe.regex="true">drops in </MotPart><MotPart gate:gateId="108492" safe.regex
="true">close to </MotPart>the hillside bank and runs deep and green. The <MotPart gate:gateId="
108495" safe.regex="true">water is </MotPart>warm too, <MotPart gate:gateId="108494" safe.regex="
true">for it </MotPart>has slipped twinkling over the yellow <MotPart gate:gateId="108497"
safe.regex="true">sands in </MotPart>the sunlight before reaching the narrow pool. On one <MotPart
gate:gateId="108496" safe.regex="true">side of </MotPart>the river the golden foothill slopes <
MotPart gate:gateId="108499" safe.regex="true">curve up </MotPart>to the strong and rocky Gabilan
mountains, <MotPart gate:gateId="108498" safe.regex="true">but on </MotPart>the valley side the <
MotPart gate:gateId="108501" safe.regex="true">water is </MotPart>lined with trees - willows
fresh and green with every spring, <MotPart gate:gateId="108500" safe.regex="true">carrying in </
MotPart>their lower leaf junctures the <MotPart gate:gateId="108503" safe.regex="true">debris of
</MotPart>the winter's flooding; and sycamores with mottled, white, recumbent limbs and branches
that arch over the pool. On the sandy bank under the trees the leaves lie deep <MotPart gate:
gateId="108502" safe.regex="true">and so </MotPart>crisp that a lizard makes a great <MotPart gate:
gateId="108505" safe.regex="true">skittering if </MotPart>he runs among them. Rabbits come <
MotPart gate:gateId="108504" safe.regex="true">out of </MotPart>the <MotPart gate:gateId="108507"
safe.regex="true">brush to </MotPart><MotPart gate:gateId="108506" safe.regex="true">sit on </
MotPart>the <MotPart gate:gateId="108509" safe.regex="true">sand in </MotPart>the evening, and
the damp flats are covered with the night <MotPart gate:gateId="108508" safe.regex="true">tracks
of </MotPart>'coons, and with the spread <MotPart gate:gateId="108511" safe.regex="true">pads of
</MotPart>dogs from the ranches, and with the split-wedge <MotPart gate:gateId="108510" safe.regex
="true">tracks of </MotPart>deer that <MotPart gate:gateId="108513" safe.regex="true">come to </
```

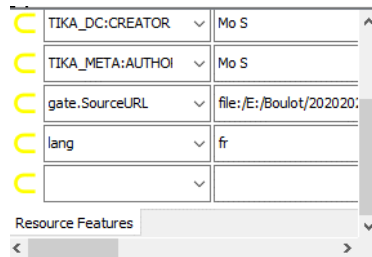
Ce qui m'a permis d'obtenir mes éléments annotés entre balises sous la forme <nomAnnotation> chaîne de caractères annotée </nomAnnotation> et donc de pouvoir y appliquer des couleurs dans mon css (background-color)

John Steinbeck's Of Mice and Men CHAPTER 1 A FEW MILES

Soledad, the Salinas River drops south of Soledad, the Salinas River drops in close to the hillside bank and runs deep and green. The water is warm too, for it has slipped twinkling over the yellow sands in the sunlight before reaching the narrow pool. On one side of the river the golden foothill slopes curve up to the strong and rocky Gabilan mountains, but on the valley side the water is lined with trees - willows fresh and green with every spring, carrying in their lower leaf junctures the debris of the winter's flooding; and sycamores with mottled, white, recumbent limbs and branches that arch over the pool. On the sandy bank under the trees the leaves lie deep and so crisp that a lizard makes a great skittering if he runs among them. Rabbits come out of the brush to sit on the sand in the evening, and the damp flats are covered with the night tracks of 'coons, and with the spread pads of dogs from the ranches, and with the split-wedge tracks of deer that come to

2. Détection de langue et Création d'une application avec des traitements conditionnels

Afin de déterminer la langue de mon document, je lui ai attribué une valeur dans l'encadré inférieur gauche.



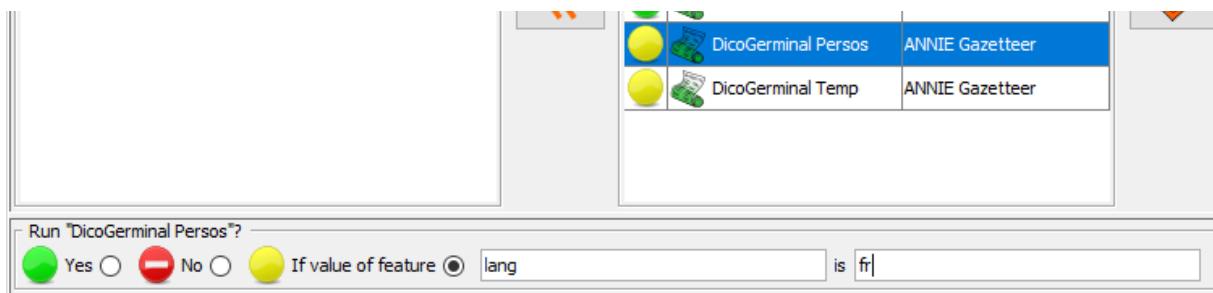
TIKA_DC:CREATOR	Mo S
TIKA_META:AUTHOI	Mo S
gate.SourceURL	file:/E:/Boulot/20202020/
lang	fr

Resource Features

C'est ce paramètre qui me permettra de créer un conditional pipeline.

Le conditional pipeline se crée exactement de la même manière qu'un pipeline classique, cependant, on peut utiliser un nombre important de Processing resources sans qu'elles soient toutes exécutées.

Pour cet exercice, j'ai donc choisi que mes dictionnaires en français ne soient exécutés que sur les textes en français tandis que les autres ressources sont exécutées sur les documents dont la langue est « en ».



DicoGerminal Persos	ANNIE Gazetteer
DicoGerminal Temp	ANNIE Gazetteer

Run "DicoGerminal Persos"?

☒ Yes ☐ No ☐ If value of feature is

Un conditional pipeline nous permet donc de ne pas multiplier les pipelines et de ne pas les manipuler trop souvent, on peut directement les exécuter sur différents corpus !

Conclusion

Cet exercice nous a permis de voir de nombreuses facettes de l'annotation de texte. Nous avons pu observer qu'on pouvait faire cela manuellement ou automatiquement mais surtout qu'il fallait anticiper le résultat de notre annotation afin de déterminer quel type d'outil utiliser.

Ainsi, pour des annotations à une ou deux occurrences, il n'y a pas besoin d'utiliser un dictionnaire ou une annotation automatique, cela va bien plus vite à la main !

Tandis que des occurrences multiples telles que les verbes, leurs flexions peuvent demander de plus amples efforts afin de déterminer comment les annoter.

On doit alors se poser des questions quant à savoir s'il vaut mieux user des outils intégrés à ANNIE, des dictionnaires ou encore des grammaires JAPE, sachant, évidemment que l'ont doit anticiper la difficulté de la tâche.

Ce projet nous a permis de mettre en exergue toutes ces questions et l'expérience nous a apportés quelques réponses.