Marina Seghier P3 – 21705105
Eirini Rozalia Ompasogkie P10 – 41006700
Doruntina Fazliu P10 – 38011906
Regina Costa P10 – 41000639

Projet d'annotation morpho-syntaxique

Dans le cadre du cours "Enrichissement de corpus", nous avons réalisé un projet d'annotation morpho-syntaxique d'un texte d'un article sur "Les personnages de Game Of Thrones et leurs homologues Pokémon" (https://www.crumpe.com/).

L'objectif principal de ce projet était avant tout de nous familiariser avec l'outil que nous avons choisi, puis de comprendre les difficultés de l'annotation, du travail en groupe et de discuter ensemble quant à la perception des annotations que nous avions sur le texte en fonction des différents points de vue.

Nous avons choisi ce texte car il y a beaucoup de noms propres de personnages dans cet article, ainsi que dans les sous-titres qui le composent.

Notre travail a consisté en une annotation manuelle individuelle, puis en une annotation automatique à l'aide du logiciel "*Visual Interactive Syntax Learning*" (https://visl.sdu.dk/visl/fr/), ainsi qu'à l'écriture d'un script en langage python composé de plusieurs fonctions.

Ce rapport se déroule en 6 étapes :

1. La méthodologie d'annotation manuelle
2. L'étiquetage automatique
3. Le calcul de précision, rappel et F-mesure
4. Les problèmes rencontrés
5. Le script (en PJ)
6. Limites et améliorations possibles de notre travail

1. La méthodologie d'annotation manuelle

Afin de réaliser l'annotation manuelle, il était indispensable de créer des règles d'annotation communes (ce que nous n'avons pas tout de suite fait).

Dans un premier temps, l'annotation manuelle a été réalisée individuellement, sans concertation. Puis dans un second temps, nous nous sommes mis d'accord sur des choix d'annotations et des changements appropriés à apporter à notre travail. Après d'éventuelles modifications, nous avons regroupé toutes les annotations individuelles, dans un fichier commun.

L'annotation manuelle consistait en l'étiquetage morpho-syntaxique. Tout d'abord nous avons segmenté nos phrases en mots. Ensuite, nous avons attribué des catégories syntaxiques (nom, verbe, adjectif...)

Comme dernière étape, nous avons ramené chaque forme rencontrée à sa forme de base (son lemme), et nous avons créé deux colonnes pour aligner notre travail (l'annotation automatique à venir et l'annotation manuelle).

Voici les POS que nous avons utilisé (à gauche, les POS prédéfinis de l'annotation automatique, et à droite, les POS que nous avons choisi pour nos annotations manuelles) :

ADJ	ADJ (Adjectif)
ADV	ADV(Adverbe)
ART	ART(Article)
PRP	PRP (Proposition)
PERS	PRON (Pronom)
PROP	PROP (Nom propre)
PRON	PRON (Pronom)
KS	KS (Conjonction)
KC	KC(Conjonction)
DET	DET (Déterminant)
N	N (Nom)
V	V (Verbe)
NUM	NUM (Numéral)
INDP NOM	PRON (Pronom)

L'annotation manuelle était fondamentale car nous avions besoin d'une annotation de référence afin d'évaluer à quel point l'annotation automatique en serait proche.

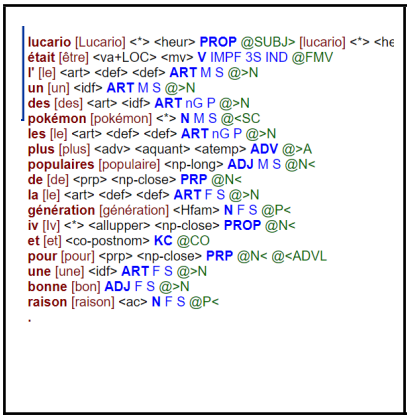
Une fois les annotations manuelles réalisées, nous avons procédé à l'annotation automatique.

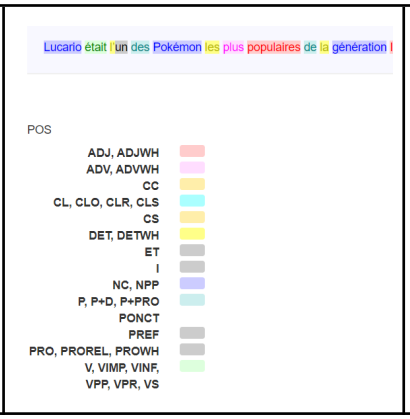
2. L'étiquetage automatique

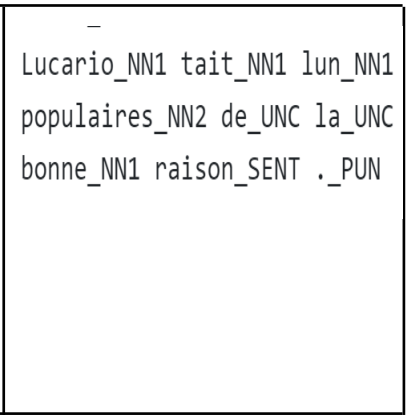
Le choix du logiciel *"Visual Interactive Syntax Learning"* a été unanime. Parmi la liste de logiciels qui nous a été fournie, nous avons considéré ce logiciel comme

étant le plus complet, intuitif et clair de tous, notamment grâce à l'explicitation des données de l'annotation réalisée.

https://visl.sdu.dk/visl/fr/parsing/automatic/complex.php	https://apps.lattice.cnrs.fr/sem/demo	http://ucrel-api.lancaster.ac.uk/cgi-bin/claws74.pl







Comparaison des annotations sur le logiciel avec nos annotations manuelles :

« Il n'y pas de meilleur jeu d'étiquettes, […] dans la pratique la plupart des jeux d'étiquettes constituent plutôt des compromis entre la finesse de la description linguistique et ce qui peut être attendu, pour des raisons pratiques, d'un système automatique d'étiquetage » Leech (1994, p.51)

Chaque annotation automatique était comparée avec l'annotation manuelle, et dans certains cas, nous avons gardé l'annotation manuelle car jugée correcte. Voici un extrait des "erreurs" (qui n'en sont pas forcément à chaque fois, mais du moins qui ne suivent pas nos choix d'annotation) que nous avons pu relever dans l'annotation automatique (à gauche) par rapport à l'annotation de référence (à droite) :

de [de] <sam-> <np-close> **PRP** @N< **les** [le] <-sam> <def> **ART** nG P @>N	**des** [de+le] **ART**
à [à] <prp> <sam-> **PRP** @<ADVL **_le** [le] <-sam> <art> <def> <-sam> <def> **ART** M S @>N	**au** [à+le] **ART**
team [team] <*> **N** M S @P< **rocket** [rocket] <*> <np-close> **N** M S @N<	**Team rocket** [team rocket] **N**

Les Pokémon [Les=Pokémon] <heur> **N** UTR S IDF NOM @P<	**les** [le] **ART** **Pokémon** [Pokémon] **N**
peut- [pouvoir] <hyfen> <*> <hyfen> <aux> **V** PR 3S IND @FAUX être [être] <va+LOC> <mv> **V** INF @ICL-AUX< **être** [être] <va+LOC> <mv> **V** INF @ICL-AUX<	**Peut-être** [peut-être] **ADV**
il [il] **PERS 3S NOM** @SUBJ>	**il** [il] **PRON**
homologues [homologue] **ADJ** M S @>N [homologue] **ADJ** nG P @>N [homologue] **N** nG P @SUBJ>	**homologues** [homologue] **N**

Comme nous pouvons le voir, l'étiqueteur pouvait par exemple :

- "décontracter" des articles comme "des" (de + le) ou "au" (à + le) ;
- couper les noms propres en plusieurs tokens ("team" et "rocket" au lieu de "Team rocket") ;
- annoter un nom avec son article comme un seul token ("Les Pokémon" au lieu de "Pokémon") ;
- annoter des adverbes comme des noms, si ces derniers étaient composés de formes dites "verbales" ("peut" et "être" pour "peut-être", car lemmatiser comme "pouvoir" et "être") ;
- annoter avec des noms de POS différents des nôtres (PERS pour PRO, restant néanmoins tous les deux la catégorie syntaxique du "pronom personnel") ;
- attribuer plusieurs lemmes et/ou plusieurs POS à un terme (par exemple ici "homologues" a été annoté comme un adjectif à 2 reprises, et comme un nom, donc 3 POS possibles).

3. Le calcul de précision, rappel et F-mesure

Afin de calculer la *précision,* le *rappel* et la *F-mesure*, nous avons utilisé trois formules (voir image ci-dessous) que nous avons rentrées dans une feuille de calcul. Ensuite, nous avons rempli ces tableaux avec 3 valeurs que nous avons comptées manuellement, à savoir : le nombre de bons résultats trouvés (dans l'annotation automatique), le nombre de résultats à trouver (dans l'annotation automatique et l'annotation manuelle), puis nous avons obtenu le résultat.

La *précision,* c'est la division du nombre d'éléments pertinents retrouvés par le nombre total d'éléments trouvés.

Le rappel, c'est le ratio du nombre d'éléments pertinents retrouvés par le nombre total d'éléments pertinents.

La *F-mesure* est la moyenne de la *précision P* et du *rappel R*.

Rappel

$$R = \frac{nb\ de\ bons\ résultats\ trouvés}{nb\ de\ résultats\ à\ trouver}$$

Précision

$$P = \frac{nb\ de\ bons\ résultats\ trouvés}{nb\ de\ résultats\ trouvés}$$

F-mesure

$$F = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

Marina :

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	249	269	Rappel	249	283
	0,9256505576			0,8798586572	
F-mesure	0,902173913				

Regina :

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	260	271	Rappel	260	332
	0,9594095941			0,7831325301	
F-mesure	0,8623548922				

Doruntina :

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	273	278	Rappel	273	316
	0,9820143885			0,8639240506	
F-mesure	0,9191919192				

Eirini :

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	274	290	Rappel	274	303
	0,9448275862			0,904290429	
F-mesure	0,9241146712				

Voici le résultat final sur nos annotations individuelles réunies :

TOTAL :

	Nb de bons résultats trouvés (automatique)	Nb de résultats à trouver (manuel)		Nb de bons résultats trouvés (automatique)	Nb de résultats trouvés (automatique)
Précision	1056	1108	Rappel	1056	1234
	0,9530685921			0,8557536467	
F-mesure	0,901793339				

(tableaux consultables dans le tableur en PJ)

En utilisant la précision, le rappel et la F-mesure, nous avons pu évaluer notre annotation. Pour comprendre nos résultats, il est nécessaire d'insister sur le fait que les valeurs de précision, rappel et F-mesure sont toujours comprises entre 0 et 1. En analysant nos données, nous constatons avoir obtenu des résultats satisfaisants car ils se rapprochent de 1.

En effet, le fichier final de nos annotations réunies présente une précision de 0,95 (soit environ 95%), un rappel de 0,85 (soit environ 85%) et une F-mesure de 0,90 (soit environ 90%). Cela signifie que notre étiqueteur trouve la quasi-totalité des éléments appropriés et ne fait que très peu d'erreurs.

Nous pouvons donc affirmer que notre système est relativement précis et pertinent.

4. Les problèmes rencontrés

Au cours de ce travail, nous avons rencontré plusieurs problèmes sur différents aspects. Principalement les problèmes d'ambiguïté de mots pluri-catégoriels. Par exemple, lorsque on a fait l'annotation manuelle nous avons dû faire des choix comme dans.

1. en [en] PREP
2. forme [forme] N

Nous avons finalement décidé de le considérer comme une préposition et un nom, au lieu de le considérer comme un adjectif.
Aussi, il a été fondamental de définir des règles de découpage et d'identification des tokens. Au début, lors de notre annotation manuelle, nous avions decidé de séparer les noms propres, mais par la suite, nous avons estimé qu'il était plus adapté de les annoter comme un seul token (exemple n°3) :

1. tywin [Tywin] PROP
2. Lannister [Lannister] PROP
3. tywin Lannister [Tywin=Lannister] PROP

5. Le script en python (en PJ)

Le script recherche et comptabilise tous les POS (Part of Speech) d'un fichier au format .txt d'annotation manuelle et automatique. Cela nous a permis de voir le nombre exact de chaque POS de chaque fichier individuellement, et de corriger nos éventuelles erreurs de comptage, avant de compiler toutes nos annotations pour n'en faire qu'une seule (manuelle et automatique).

Ce script est composé de plusieurs fonctions :
- une première **count_pos()** (faite de 2 manières différentes, l'autre étant **recherche_pos()**) qui calcule le nombre total de POS d'un fichier donné ;
- une deuxième **PRF()** qui calcule la précision, le rappel et la F-mesure (avec un appel à la première fonction) ;
- une troisième **noms_propres()** qui calcule le nombre de noms propres d'un fichier donné.

Cette dernière nous permettait d'observer l'écart entre les 2 annotations et voir s'il était significatif concernant les noms propres (mais la fonction peut s'appliquer à n'importe quel POS en modifiant la regex). Il nous a semblé intéressant d'observer les noms propres spécifiquement puisque l'étiqueteur les découpait soit en 2 voire 3, soit les gardait en entier.

Output du script python :

```
Nombre de POS comptabilisés dans :
L'annotation automatique : 1234
L'annotation manuelle : 1108

La précision est de : 0.9530685920577617
Le rappel est de : 0.8557536466774717
La F-mesure est de : 0.9017933390264731

Nombre de noms propres comptabilisés dans :
L'annotation automatique : 119
L'annotation manuelle : 108
```

Nous retrouvons les valeurs présentent dans le tableur (1234, le nombre de résultats trouvés automatiquement et 1108, le nombre de résultats à trouver manuellement).
Les calculs de précision, de rappel et de F-mesure sont également corrects. Et dans l'ensemble, les noms propres ont correctement été annotés (119 trouvés dans l'annotation automatique contre 108 dans l'annotation manuelle).

6. Conclusion

En conclusion, ce projet nous a permis de découvrir les difficultés de l'annotation manuelle, automatique et du travail en groupe par rapport aux différentes perceptions des unes et des autres. Nous avons pu découvrir tous les aspects de l'annotation linguistique, en procédant à l'annotation individuelle du texte que nous avons choisi.

Une limite que nous avons pu rencontrer peut être l'incapacité de la machine à tout annoter correctement et uniformément, notamment quant aux noms propres et aux articles. De plus, le site "*Visual Interactive Syntax Learning*" ne nous donnait pas la possibilité de télécharger directement les annotations automatiques.

Une autre limite s'est posée au niveau de l'écriture du script, quant au repérage du nombre de bons résultats trouvés dans l'annotation automatique : en effet, cette valeur a été rentrée manuellement dans le script afin de calculer la précision, le rappel et la F-mesure, car la limite de temps ne nous permettait pas d'écrire un programme qui puisse repérer automatiquement quand un résultat est bon ou non dans l'annotation automatique.

Donc concernant le site, une suggestion d'amélioration possible serait de nous donner la possibilité de télécharger nos annotations automatiques et de sélectionner le format souhaité. Concernant notre travail, nous pourrions utiliser plusieurs logiciels et comparer les résultats. Ce serait vraiment intéressant et cela nous permettrait de juger quel est le meilleur annotateur automatique.