IFT6285 – Devoir 6

Analyse Syntaxique avec SpaCy

Maxime MONRAT

Juan Felipe DURAN

Université de Montréal

Automne 2021

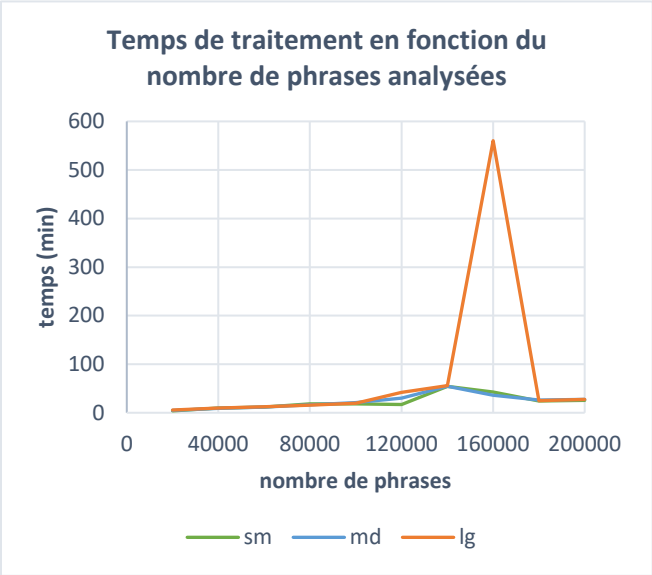
Objectifs

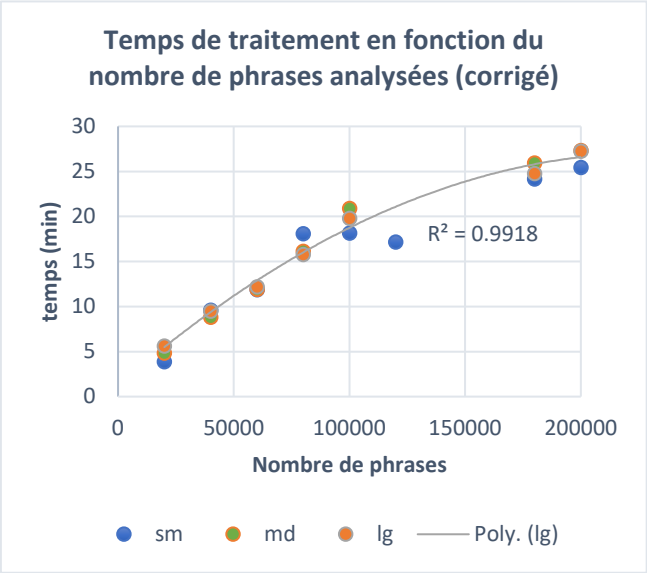
Dans ce devoir nous explorons les capacités d'analyse grammaticale offertes par la librairie SpaCy. Nous observerons les performances de celle-ci pour les analyses en dépendance à l'aide du corpus *1 Billion Word* et de trois tailles de modèles de langues incorporés à SpaCy, entraînés sur le Onto Note 5[1].

1. Temps d'analyse

On remarque que le temps d'analyse augmente avec le nombre de phrases du jeu de données, mais pas de manière significative en fonction de la largeur du modèle.

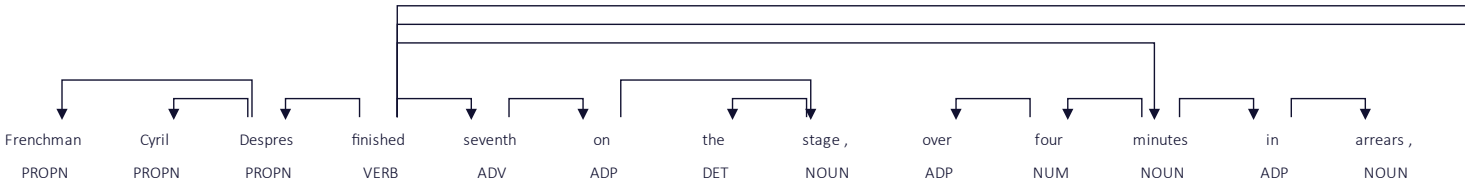
Nous devons cependant préciser que ces données sont à prendre avec prudence : le pic dans le temps de traitement pour les trois modèles entre 140 000 et 160 000 phrases semble anormal et peut être du à un processus en arrière plan sur notre machine.



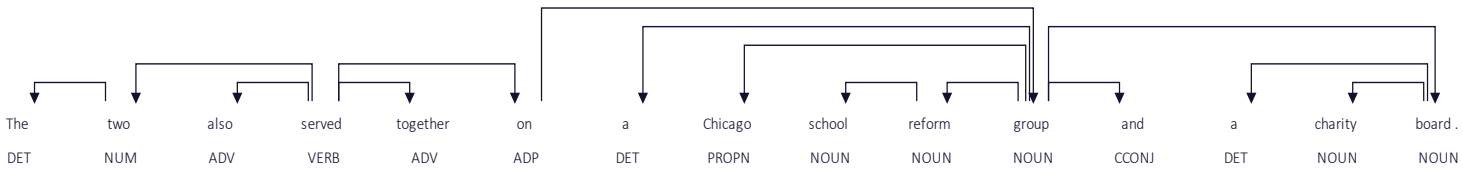


2. Analyse syntaxique en dépendance : analyses fautives

Voici 5 analyses en dépendance effectuées avec le petit modèle qui comportent des fautes :

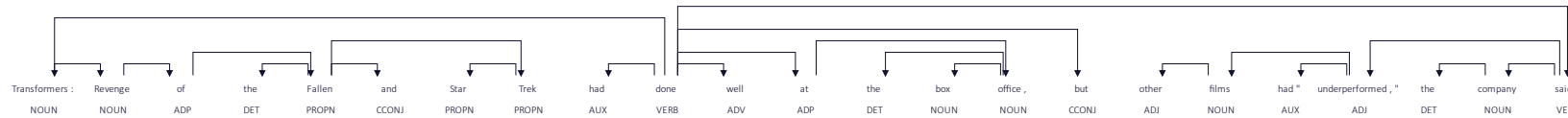


Dans ce premier exemple, le mot *Frenchman* a été analysé comme un nom propre plutôt qu'un adjectif.

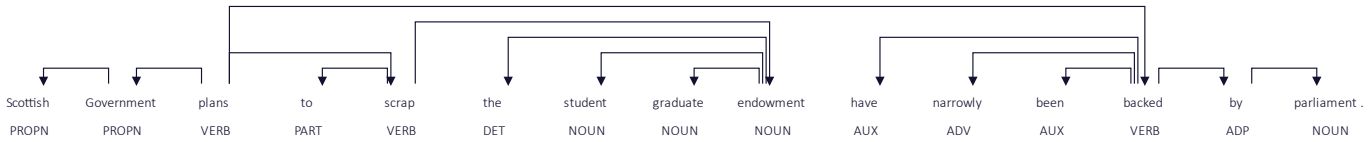


Ce second exemple montre une confusion dans les groupes nominaux *Chicago school* et *reform group*

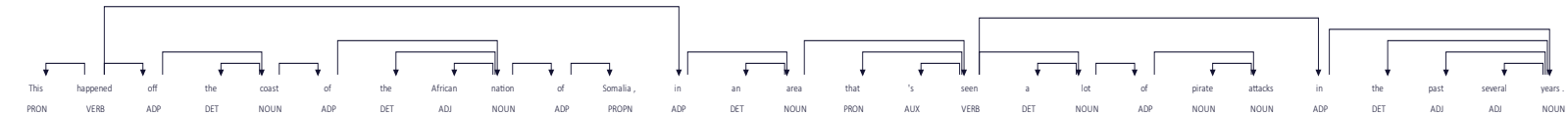
[1] OntoNotes 5 (Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, Ann Houston)



Cet exemple-ci montre une confusion dans la dépendance (non-visible ici) : le modèle n'a pas réussi à déterminer le sujet du groupe verbal *had done*, et a alors créé un lien non spécifié (*dep*) avec le sujet *Transformer*



Ce quatrième exemple montre une erreur dans l'étiquetage morpho-syntaxique : le mot *plans* est ici analysé comme un verbe, tandis qu'il s'agit d'un nom dans ce contexte-ci.



Dans ce dernier exemple nous observons également une erreur dans l'étiquetage morpho-syntaxique : *lot* est ici décrit comme un nom plutôt que comme un adverbe.

En analysant ces mêmes phrases avec le modèle large, on peut faire les observations suivantes :

- L'analyse morpho-syntaxique est sensiblement meilleure (*Frenchman* n'est plus un nom propre, mais bien un adjectif), mais reste confuse pour les mots à sens multiples tels que *plans* et *lot*
- Les dépendances sont mieux identifiées : le groupe verbal *had done* de notre troisième exemple possède bien un sujet.

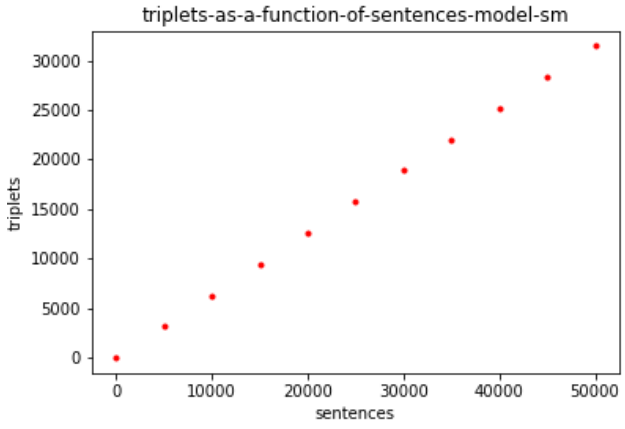
3. Analyse de sous-ensembles syntaxiques

Pour cette tâche, il fallait extraire des analyses produites tous les triplets (sujet, verbe, objet) correspondants aux lemmes des nœuds (s, v, o).

Nous avons acquis les résultats suivants pour chaque modèle :

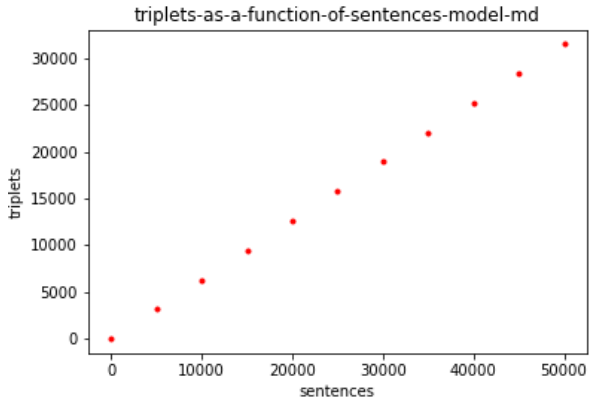
Modele_sm:

sentences	triplets
0	0
5000	3172
10000	6250
15000	9385
20000	12539
25000	15704
30000	18900
35000	22028
40000	25142
45000	28323
50000	31447



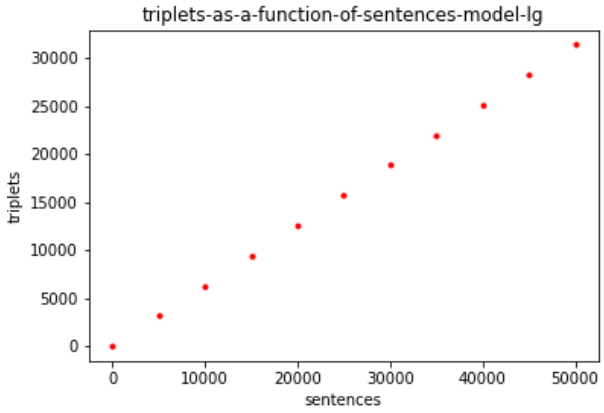
Modele_md

sentences	triplets
0	0
5000	3181
10000	6289
15000	9455
20000	12618
25000	15779
30000	18993
35000	22116
40000	25224
45000	28396
50000	31539



Modele_lg

sentences	triplets
0	0
5000	3164
10000	6258
15000	9404
20000	12546
25000	15695
30000	18875
35000	21993
40000	25091
45000	28256
50000	31386



Les trois modèles sont similaires, mais le nombre de triplets acquis par chaque modèle est différent. On remarque, par exemple, que le **modele_sm** a **31447** triplets pour 50000 phrases, le **modele_md** a **31539** triplets pour 50000 phrases et le **modele_lg** a **31386** triplets pour 50000 phrases.

4. Informations intéressantes extraites des triplets

Après avoir analysé les diffèrent triplets, nous avons remarqué quelques informations intéressantes.

Quand **couple** est le sujet, le verbe et l'objet ont beaucoup d'information relié à leur relation ou aux enfants. Par example, il y a *couple made official, couple exchanged vows, couple spending night, couple married relationship, couple gave children, couple looking life, couple started plan*, etc.

Quand **président** est le sujet, le verbe et l'objet ont beaucoup de lien avec les relations internationales, la diplomatie et les décisions : *president pushed congress, president declare diplomacy, president inked agreement, president declared era, president says Lebanon, president comitted missions, president faces turmoil, president think options,* etc.

Quand **firefighters** est le sujet, les phrases sont beaucoup plus aux alentours des thèmes reliés au feu et à l'évacuation: *firefighters wearing apparatus, firefighters rescued mother, firefighters ordered evacuation, firefighters battled fire, firefighters divert flames, firefighters decided hydrant, firefighters lost lives, firefighters tackling blaze, firefighters used pumps,* etc.

On observe un contraste lorsque le sujet est **parents**, comparé a lorsque le sujet est **students**.

Les thèmes reliés aux parents sont les responsabilités, les enfants, la vie, tandis que pour les étudiants, les thèmes tournent autour de l'éducation et les manifestations.

> **Parents** : *parents choosing schools, parents accepted money, parents adopt budget, parents held children, parents named girl, parents give toddlers, parents drop kids, parents suspected son, parents escorting students,* etc.
>
> **Students**: *students following articles, students aim university, students earn diplomas, students taking qualifications, students taking classes, students attend elementary, students challenging courses, students started strike, students finished projects, students protesting selection, students prepare candidates,* etc.

On a aussi remarqué un contraste entre le sujet est **men**, et le sujet **women**.

Par exemple, les thèmes de **women** sont plus reliés à la maternité, l'apparence et l'argent, tandis que ceux des **men** sont reliés à des aspects plutôt actifs ou agressifs.

> **Women** : *women accused children, women feel children, women enjoy rights, women prefer men, women sold swimsuits, women taught mother, women evening dress, women gathered money, women stopped estrogen, women expect mothering, women formed friendships, women need shoes,* etc.
>
> **Men**: *men enter mines, men patrolled border, men said war, men hurled insults, men threatened staff, men tipped killed, men accused man, men passed suspects, men harpoon whales, men pleaded explosions, men lost lives, men killed planes, men landed helicopters, men debated what,* etc.

Le fait de pouvoir utiliser ces triplets aussi facilement avec SpaCy est un réel atout. Observer les contrastes entre les champs lexicaux reliés aux différents types dans un corpus de textes permet de déterminer plus facilement les différents biais dans le vocabulaire, et éventuellement faciliter leurs corrections dans les étapes de plongement de mots ou de documents