Aix-Marseille University

Projet de Recherche

Traitement des Langues naturelles et Linguistiques

Mohamed Ouerfelli *et Achraf* Hamane

Encadré par
M.Alexis NASR

5 février 2020

CHAPITRE 1

Présentation générale

1.1 Introduction

Chaque année, la Conférence sur le traitement automatique des langages naturels et linguistiques propose une tâche partagée dans laquelle les participants entraînent et testent leur modèles sur les mêmes ensembles de données pour faciliter la comparaison.

1.2 Vocabulaire

Définissons un vocabulaire qui rendra le travail plus clair :

— La sémantique est le domaine linguistique et philosophique qui étudie le sens et l'interprétation. Il faut beaucoup de liens entre les mots pour comprendre la phrase et analyser les changements de sens. En programmation, la sémantique est le résultat attendu d'un programme.

— La syntaxe est le champ linguistique de la grammaire. C'est l'étude des règles pour les modèles de mots dans les phrases. Connues également en programmation, les erreurs de syntaxe conduisent souvent à des erreurs, car les règles sont souvent beaucoup plus strictes que dans le langage oral.

1.3 Qu'est-ce qu'un analyseur de dépendance ?

Un arbre de dépendance est une structure qui peut être définie comme un graphe orienté, avec |V| nœuds (sommets), correspondant aux mots, et |A| Arcs, correspondant aux dépendances syntaxiques entre eux.
Nous pouvons également vouloir attribuer des étiquettes à des dépendances, appelées relations. Ces relations fournissent des détails sur le type de dépendance.

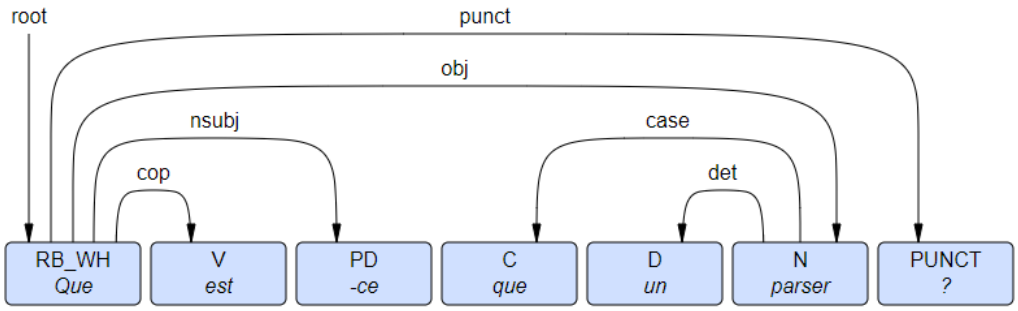


FIGURE 1.1 – Dépendence dans une phrase

Pour un arc de h vers d, h représente la "head" et d le "dépendant". La "head" est le nœud le plus important d'une phrase, tandis que la racine est le nœud le plus important de la phrase : la "head" de tous les autres nœuds. Un analyseur de dépendance transforme simplement une phrase en un arbre de dépendance.

1.4 Fonctionnement

Les analyseurs ne sont pas du tout efficaces, car les langues sont très complexes et changent avec le temps. Tout petit changement dans la langue entraînerait des changements considérables dans l'analyseur.
Plusieurs étapes sont nécessaires pour créer un analyseur de dépendance. Nos entrées sont les mots de la phrase avec leurs propriétés (index, balise Part of Speech, Lemma...) ; ensuite, nous devons chercher les caractéristiques de tous les arcs possibles de la phrase. Grâce à ces fonctionnalités, nous calculons un score pour chaque possibilité.

1.5 Métriques : comment reconnaître un bon analyseur ?

Un analyseur de dépendance précis reconnaît bien les dépendances et les relations entre les mots. Deux métriques (scores) sont utiles pour cela :
- **Score UAS (Unlabeled Attachment Score)**, qui correspond au nombre de dépendances correctement prédites sur le nombre de possibilités ;
-**Score LAS (Labeled Attachment Score)**, qui correspond au nombre de dépendances et de relations correctement prédites sur le nombre de possibilités. Le LAS est toujours inférieur ou égal au UAS, car une dépendance incorrecte conduit à un UAS et à un LAS sous-optimaux, alors qu'une relation (ou une étiquette) incorrecte conduit uniquement à une diminution du LAS.

1.6 Format Conllu

Les annotations sont codées dans des fichiers de texte brut avec trois types de lignes : Les phrases consistent en une ou plusieurs lignes de mots, et les lignes de mots contiennent les champs suivants :

- ID : index de mots
- FORME : forme de mot ou symbole de ponctuation.
- LEMMA : Lemma ou tige de forme verbale.
- UPOS : balise de partie de parole universelle.
- XPOS : balise de partie de la parole spécifique à la langue.
- FEATS : liste des caractéristiques morphologiques de l'inventaire des caractéristiques universelles ou d'une extension définie spécifique à la langue.
- DEPREL : Relation de dépendance universelle avec HEAD.

1.7 Caractéristiques et score

1.7.1 Features

Avec les attributs vus dans la partie précédente, on extrait les variables explicatives et nous entraînons un modèle de régression multiple qui nous permet d'expliciter la corrélation entre ces features et le score (label).

L	LAS	UAS	L	LAS	UAS	L	LAS	UAS
hi	79.47	86.80	ru	69.70	73.85	sl	63.47	71.78
it	78.38	82.15	da	68.12	74.18	hr	63.58	72.10
ur	76.33	83.55	id	67.05	72.21	cs	63.84	72.45
pl	76.18	84.41	en	67.18	74.39	lv	62.30	69.83
ja	75.74	85.60	es	66.93	74.52	hu	62.73	68.86
no	73.25	78.91	uk	65.85	74.19	fi	62.77	70.83
bg	73.40	82.36	ro	65.13	72.53	zh	59.91	65.15
el	72.55	78.52	ga	65.13	74.02	vi	59.77	62.68
ca	72.06	79.70	fa	65.22	73.42	eu	58.80	68.78
sv	71.10	77.36	he	64.68	72.34	nl	57.44	68.43
fr	71.36	77.02	et	64.76	75.40	ko	53.12	63.21
pt	70.73	76.95	ar	64.28	71.65	tr	47.28	55.20

TABLE 1 – Labeled Accuracy Score (LAS) et Unlabeled Accuracy Score (UAS) pour 36 langues différentes dans des conditions d'apprentissage proches.

1.7.2 Score R^2

Pour mesurer la proportion de variabilité de y qui est expliquée par le modèle, on utilise le **coefficient** R^2. Plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données.

$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$

avec $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y_n})^2$, $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ et $SCT = \sum_{i=1}^{n}(y_i - \bar{y_n})^2$

CHAPITRE 2

Extraction des variables explicatives

2.1 Introduction

Dans cette partie, nous allons analyser plusieurs variables explicatives et leur coefficient R^2 associé. Une variable prise seule, dans une régression linéaire simple, peut parfois donner un score R^2 faible. Mais lorsqu'on l'assemble aux autres variables lors de la régression multiple, le score R^2 lié à toutes les variables peut nettement s'améliorer.

Pour pouvoir étudier nos données, nous allons séparer notre étude en différentes parties :

— Visualisation des données

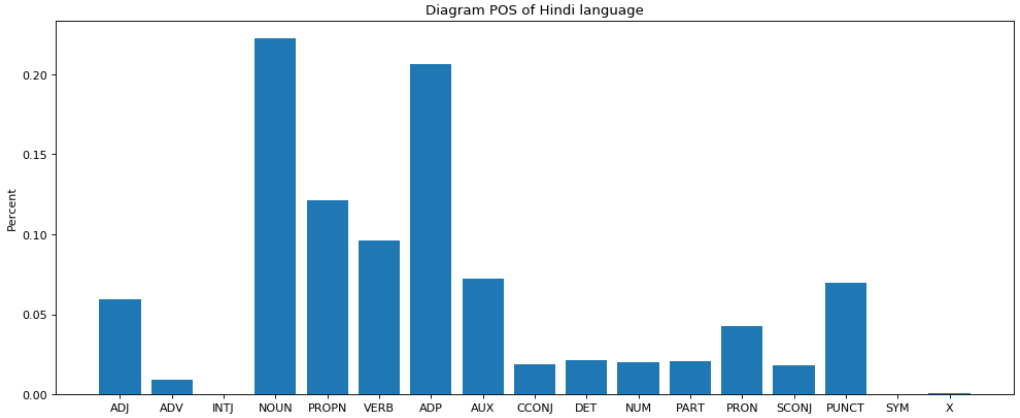
— Variables explicatives

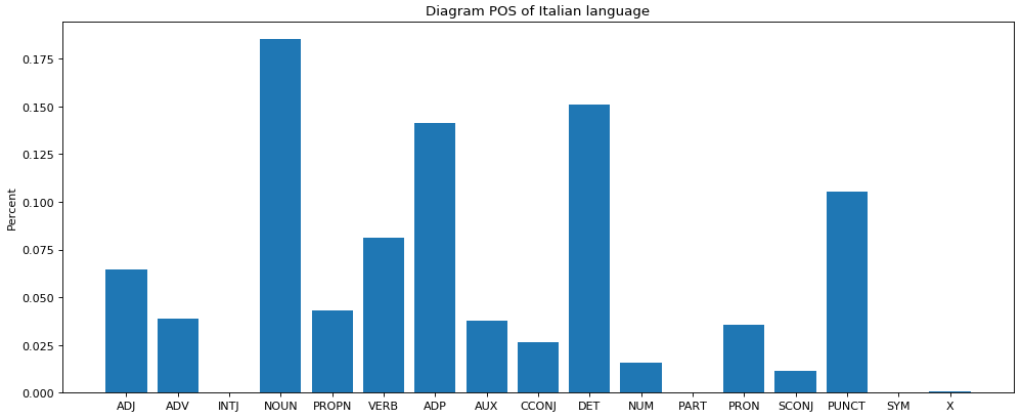
— Régression multiple

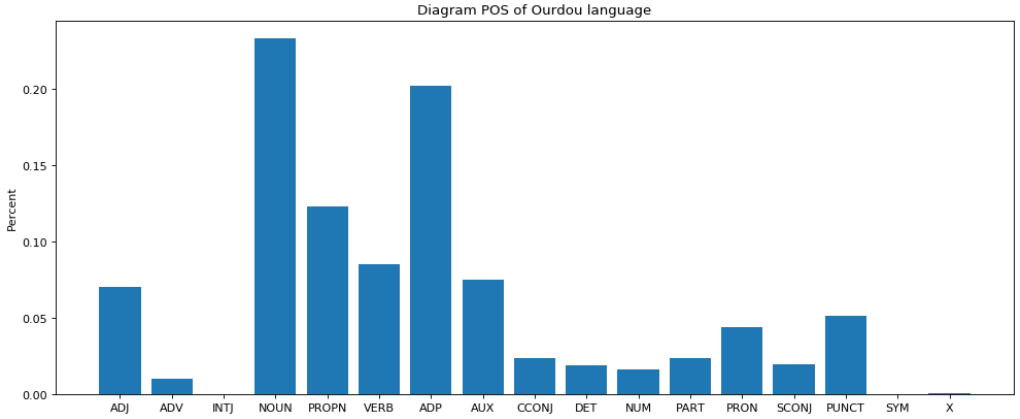
— Amélioration de l'analyseur

2.2 Visualisations des données

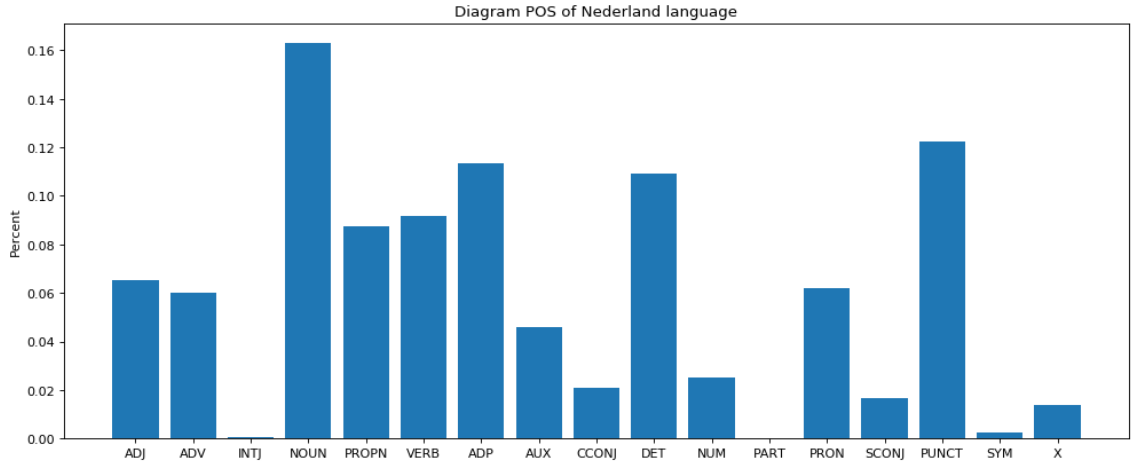
On s'intéresse seulement aux Part Of Speech (POS) en affichant dans un diagramme à barre la part d'une POS par rapport à toutes les autres.
Tout d'abord, on a choisi de visualiser les POS des 3 langues avec les meilleurs scores LAS :

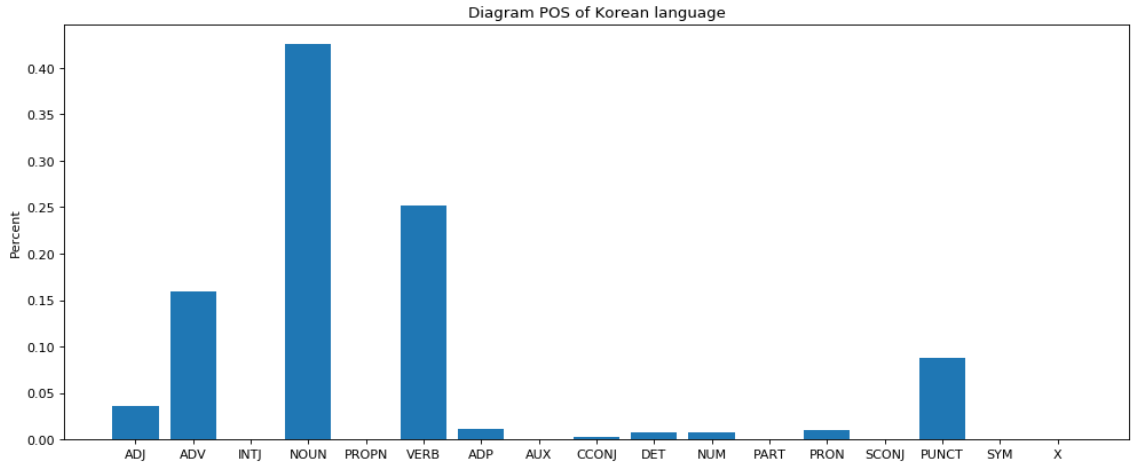


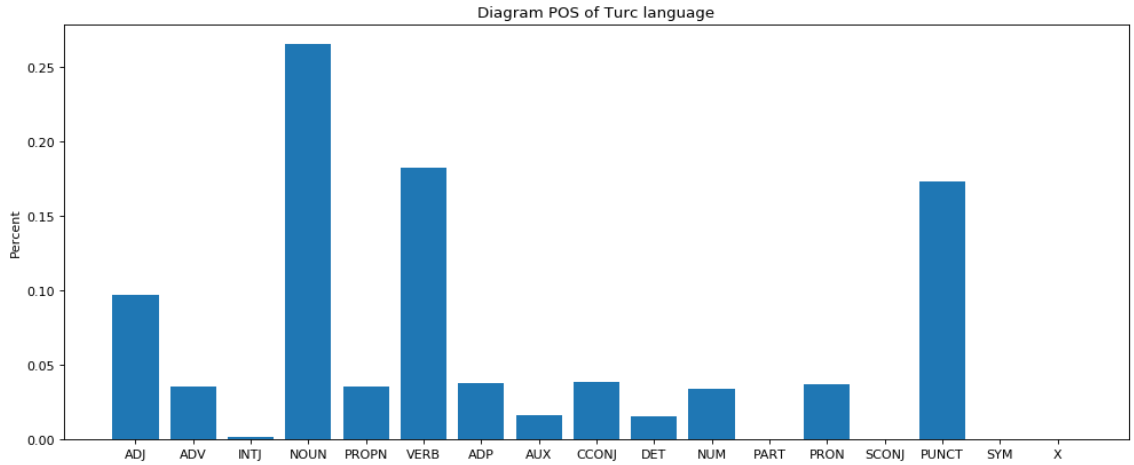




On observe ensuite les POS des 3 langues avec les scores LAS les plus faibles :







Remarques : On remarque, par exemple, que la POS *Adposition* représente un pourcentage par rapport à tous les POS plutôt élevé pour le hindi, l'italien et le ourdou qui sont les 3 langues qui ont les meilleurs scores LAS. Alors que pour le

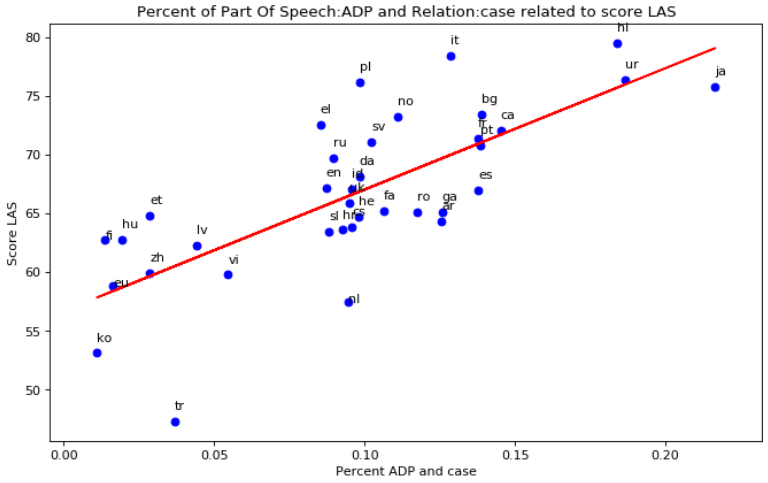
coréen et le turc, qui sont les 2 langues avec les scores LAS les plus faibles, la part de *Adposition* est très basse par rapport aux autres POS.

Par ailleurs, on comprend facilement que la *punctuation* ne sera certainement pas une bonne variable explicative, si on se fie à ces diagrammes. En effet, on a une grande variation de ce dernier, d'une langue à une autre, quelque soit le score LAS de la langue en question.

2.3 Variables explicatives

2.3.1 Pourcentage de Part Of Speech & Relations

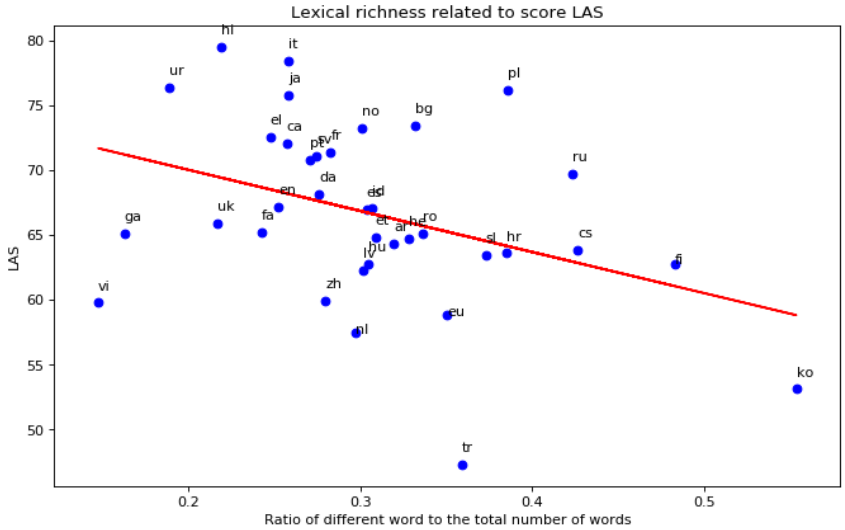
Certaines POS et relations, sont plus présentes pour certaines langues plutôt que d'autres. Effectivement, si on calcule le pourcentage de la POS Adposition, et de la Relation case on obtient les obversations suivantes :



Avec cette variable explicative on a environ 0.55 comme score R^2.
C'est la variable explicative avec le coefficient R^2 le plus élevé que l'on a réussi à obtenir. On remarque, en effet, que certaines langues comme le ourdou, le japonais ou le hindi ont un score LAS très élevé et un pourcentage d'Adposition et de case élevé aussi. Alors que le turc ou le coréen ont des scores LAS faibles et un pourcentage d'Adposition et de case faible.

2.3.2 Richesse Lexicale

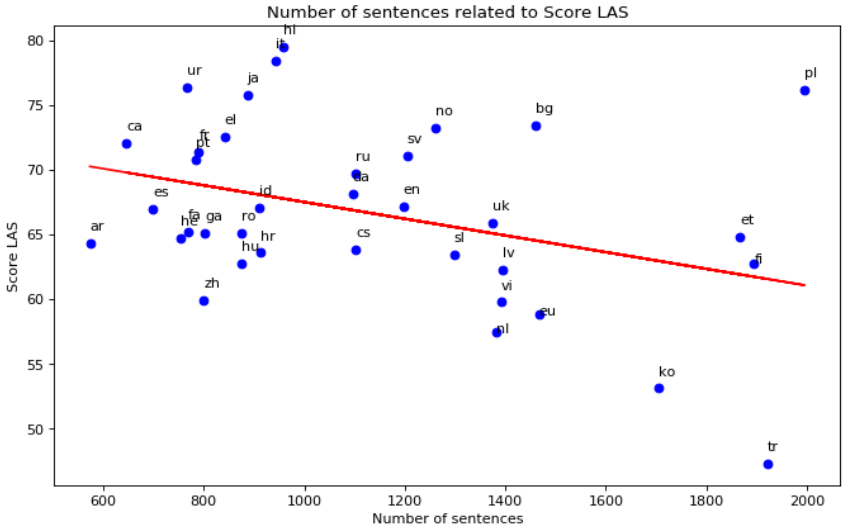
Une façon de déterminer si une langue est riche lexicalement ou non est d'analyser le nombre de mots différents par rapport au nombre total de mots du corpus. En appliquant un algorithme pour toutes les langues de notre étude, on peut en obtenir le grapique suivant :



Ici on obtient un score R^2 d'environ 0.14.
Cette variable explicative prise seule n'explique pas vraiment pourquoi certaines langues ont un score LAS meilleur que pour d'autres langues.

2.3.3 Nombre de phrases

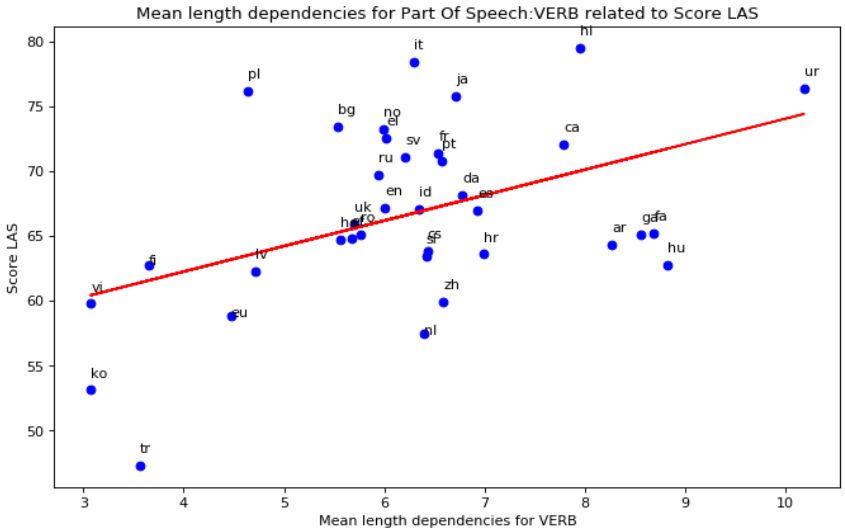
Pour cette variable, nous avons simplement calculé le nombre de phrases pour chaque langue.



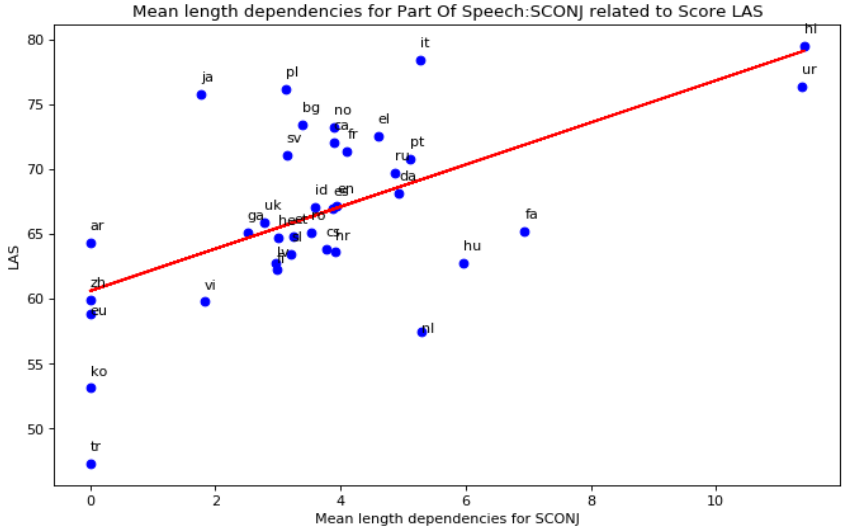
Le score R^2 est égal à 0.1327.
On remarque, qu'à certaines exception près, que moins il y a de phrases dans le corpus, plus le score LAS est elevé.

2.3.4 Longueur moyenne des dépendances

Ici, nous avons étudié la longueur moyenne des dépendances pour chaque langue, Pour obtenir ces résultats, on fait une moyenne portant sur une simple soustraction en valeur absolue entre l'identifiant du gouverneur et l'identifiant du dépendant, pour tous les mots du fichier.
Et nous nous sommes intéressés à chaque fois à cette moyenne pour une seule POS, ou une seule Relation.



Remarque : Les langues qui ont une longueur moyenne des dépendances liée aux POS VERB très élevée ont des scores LAS élevé comme le ourdou par exemple. Alors qu'une langue comme le turc a un score LAS faible ainsi qu'une longueur moyenne des dépendances liée aux verbes faible.
La longueur moyenne des dépendances liée à POS VERB nous permet d'obtenir un score $R^2 = \mathbf{0.20}$.



Remarque : Ce graphique, nous permet de visualiser le fait que des langues avec un score LAS élevé comme le hindi ou le ourdou ont une longueur moyenne des dépendances liée à SCONJ très élevée. En opposition, à des langues comme le turc et le coréen qui ont des scores LAS et une longueur moyenne des dépendances liée à SCONJ très bas. Le score R^2 obtenu ici est de **0.35**.

On applique par la suite le même algorithme pour d'autres Relation et Part Of Speech, et on a le tableau de score R^2 suivant :

	case	amod	flat	obj	clf	ADV
R^2	0.0056	0.0098	0.0873	0.0498	0.0277	0.0012

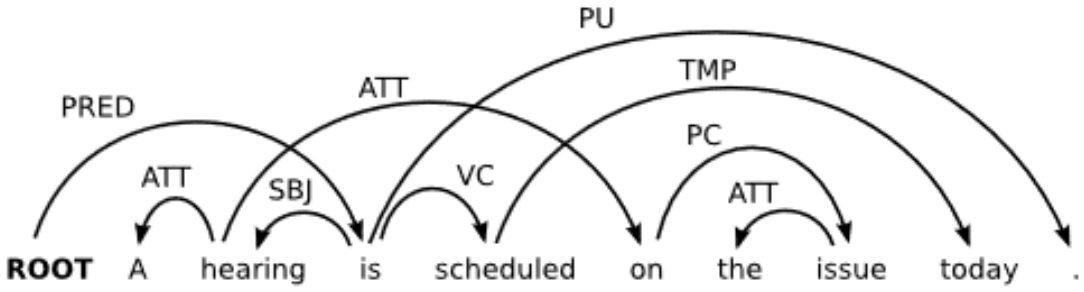
	DET	cc	ADJ	mark	NOUN
R^2	0.0411	0.1057	0.0567	**0.1302**	**0.1451**

Commentaires :

- Parmi toutes les possibilités, c'est la longueur moyenne des dépendances liée à la conjonction de Subordination qui nous permet d'obtenir le meilleur score $R^2 = \mathbf{0.35}$.
- La plupart des scores sont assez faibles, mais lorsqu'on les couple ensemble, cela permet d'avoir un score R^2 lié à la régression multiple important.

2.3.5 Taux de non-projectivité

Dans cette partie, on s'intéresse au taux de phrases non-projective. Une phrase est non-projective si on a un croisement entre 2 flèches de relations. Voici un exemple d'une phrase en anglais non-projective :



En créant un algorithme qui calcule le taux de non-projectivité, on obtient un score R^2 de 0.1043.

2.3.6 Régression linéaire multiple

Avec les variables explicatives trouvées dans la partie précédente, nous avons entraîné un modèle de régression linéaire multiple qui nous permet d'expliquer la corrélation entre ces features et le score LAS. Au final, lorsque l'on assemble ces **17** variables et en appliquant cette régression, on obtient un score R^2 de **0.911**.

```
regressor = LinearRegression()
regressor.fit(X, y)     # X contient toutes les variables explicatives
y_pred = regressor.predict(X)
from sklearn.metrics import r2_score
print( "## r2 score :##")
print(r2_score(y, y_pred, multioutput='uniform_average'))
```

```
## r2 score :##
0.9110033085987238
```

Mis en évidence de l'explicabilité de certaines variables :

Il est aussi possible d'appliquer la régression multiple à un nombre plus petit de variables, puisque certaines variables ont un rôle beaucoup plus important que d'autres dans l'explication du score LAS. En effet, on peut faire les combinaisons suivantes avec seulement les variables les plus explicatives qui nous permettent d'avoir un plutôt bon score.

La régression linéaire multiple peut alors être appliquée sur seulement :

- **deux variables :** La variable explicative "Pourcentage de Part Of Speech Relations" vue dans la sous-section 2.3.1 et la variable "Longueur moyenne des dépendances de SCONJ" vue en 2.3.4.
  Dans ce cas, on obtient un score R^2 élevé de 0.61.

- **trois variables :** En reprenant les deux variables, observées précédemment, et en y ajoutant la variable explicative "Taux de non-projectivité" vue dans la sous-section 2.3.5.
  Dans ce cas, on obtient un score R^2 encore meilleur de 0.66.

2.4 Amélioration de l'analyseur MACAON

Une amélioration possible de l'analyseur mis à notre disposition serait d'ajouter les variables explicatives les plus importantes choisies dans la section précédente au fichier fm de l'analyseur. Cela permettrait d'améliorer les scores obtenus par MACAON.

CHAPITRE 3

Conclusion

La complexité et la variabilité des langues au cours du temps rendent compliqué la tâche d'un analyseur. Un changement minime dans une langue peut avoir d'importantes conséquences sur l'analyseur.
L'apprentissage automatique et le Deep Learning permettraient de développer des analyseurs syntaxiques empiriques, basés sur les données, qui s'adapteraient plus facilement à ce genre de problèmes.