Treetagger pour l'étiquetage des parties du discours du mot *that*

Kpodjro Kpatoukpa
kpatoukpakpodjro@gmail.com
Lyse Priscille Ngasseu Ndifo
lysendifo8@gmail.com
Serigne Bassirou Ndiaye
dstndiaye@gmail.com
Orchelle Patricia Welehela Taweuteu
patriciawelehela@gmail.com

Abstract

Cette étude explore l'utilisation de TreeTagger pour étiqueter les parties du discours du mot *that*. Dans un premier temps, nous avons présenté les résultats obtenus (métriques d'évaluation) pour la catégorisation de *that*, en utilisant les modèles pré-entraînés BNC et Penn, suivie d'une comparaison entre les deux.

Ensuite, pour améliorer la précision de l'étiquetage, nous avons réentraîné TreeTagger avec un jeu d'étiquettes provenant du corpus Brown. Il inclut des étiquettes telles que WPR pour les pronoms relatifs, CST pour les conjonctions (propositions nominales), CJT pour les conjonctions (propositions verbales), DT pour les déterminants, et RB pour les adverbes.

D'autres évaluations et comparaisons ont été faites avec différentes modèles, comme Stanza et UDpipe.

Cette étude propose des méthodes pour améliorer l'étiquetage des parties du discours de *that* et offrir ainsi un outil plus précis pour une analyse linguistique centrée sur les différentes fonctions de *that*.

1 Crédits

Nous tenons à remercier Monsieur Nicolas Ballier Prof of English Linguistics at l'Université de Paris Cité pour ses conseils et son soutien tout au long du cours et de la réalisation du projet.

Nous souhaitons également exprimer notre reconnaissance envers les contributeurs du corpus **Brown**, qui a été utilisé pour l'entraînement de notre modèle.

Les personnes ayant contribué à son développement sont : Henry Kucera (Linguiste et Informaticien), W. Nelson Francis (Linguiste) et d'autres chercheurs de l'Université Brown.

Ce corpus est accessible via le Linguistic Data Consortium (LDC) et est aussi disponible sous une version annotée dans le NLTK.

2 Introduction

L'étiquetage des parties du discours est une tâche essentielle en traitement automatique des langues, car il permet d'assigner des catégories grammaticales aux mots d'un texte. TreeTagger est un outil utilisé pour cette tâche, mais il peut avoir des difficultés à distinguer les différentes utilisations d'un mot.

Dans cette étude, nous nous intéressons à la distinction entre le *that* relatif et le *that* des complétives nominales. Notre objectif est de réentraîné TreeTagger avec un jeu d'étiquettes provenant du corpus Brown, afin d'améliorer la précision de l'étiquetage de ces différentes réalisations de *that*.

Pour ce faire, la précision des modèles pré-entraînés BNC et Penn sera d'abord évaluée en termes de catégorisation de *that*, et servira de référence pour l'amélioration des performances. Ensuite, TreeTagger sera réentraîné et d'autres évaluations seront faites en utilisant différentes modèles telles que Stanza et UDpipe. L'impact de la taille des données d'entraînement sur la précision de l'étiquetage sera également analysé. De plus, des techniques de fine-tuning et d'autres approches seront testées pour améliorer les performances des modèles.

3 Description des données

3.1 Données de test

Les données de test utilisées pour évaluer les modèles pré-entraînés Penn Treebank et BNC sont constituées de fichiers *.txt* contenant 100 phrases chacune avec différentes réalisations de *that*. Ces

fichiers sont organisés par catégories : déterminant, conjonction(pour les verbes, CJT), pronom et adverbe. Les données ont été récupérées à partir des fichiers fournis dans le cadre du cours de Data Science avancée. D'autre part, des données synthétiques ont été aussi utilisées pour tester l'étiquetage (après le réentrainement de treetagger) de *that* en tant que conjonction pour les noms (CST).

La figure ci-dessous illustre le nombre de phrases et les occurrences totales du mot *that* dans différentes catégories linguistiques.

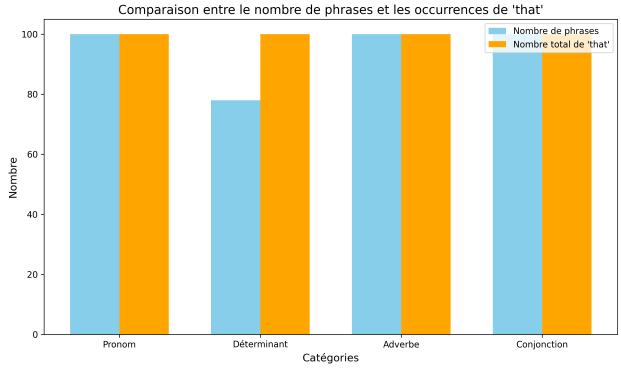

Figure 1: Nombre de phrases et les occurrences de *that* par catégorie

3.2 Données d'entraînement - Corpus Brown

Il a été créé à l'Université de Brown en 1961. Il rassemble des textes américains et est divisé en 15 genres textuels différents, chacun représentant un type spécifique de texte. Il comprend des genres variés tels que la fiction, le journalisme, les essais et d'autres formes littéraires. Ce corpus contient environ un million de mots répartis sur plusieurs centaines d'extraits, chacun présentant un style et un contexte différents.

Le corpus Brown a été choisi pour le réentraînement de TreeTagger en raison de sa diversité, couvrant une large gamme de genres textuels et offrant ainsi une meilleure représentation de la langue anglaise. Son statut de corpus de référence facilite l'évaluation et la validation des modèles, permettant des comparaisons fiables. De plus, son étiquetage grammatical, réalisé manuellement par des experts, garantit une grande précision et une qualité élevée des annotations, ce qui en fait un choix pertinent pour l'amélioration des performances de TreeTagger.

Figure 2: Word cloud des mots les plus fréquents

L'analyse du nuage de mots généré à partir du corpus Brown montre que, les mots qui dominent sont *the*, *and*, *of*, *in*, *to* et *a*. En comparaison, le mot *that* apparaît avec une fréquence relativement moindre. Néanmoins, son utilisation demeure significative puisqu'il intervient dans des constructions syntaxiques et stylistiques spécifiques. Ainsi, même s'il est moins répandu que les mots dominants, *that* continue de jouer un rôle particulier dans l'organisation du discours au sein du corpus Brown.

Les phrases contenant le mot *that* ont été extraites et annotées dans le corpus Brown. Le tagset C8 a ensuite été adapté pour cibler spécifiquement les différentes utilisations de *that*. Par la suite, le mapping des tags d'origine a été appliqué vers ce nouveau tagset, et un corpus d'entraînement au format TreeTagger a été généré.

Ainsi, lorsqu'on examine la distribution du mot *that* selon les catégories linguistiques, il apparaît qu'il est utilisé de manière polyvalente :

- CST (Conjonction pour les noms) : Avec 3 831 occurrences, elle comprend les cas où *that* est utilisé pour introduire une proposition subordonnée relative ou un complément, comme dans l'expression *the fact that*;
- WPR (Pronom relatif) : Avec 1 662 occurrences, désigne une entité mentionnée avant, souvent dans des propositions relatives. Par exemple, *The book that I read.*;
- DT (Déterminant) : Avec 2 272 occurrences, est utilisé pour désigner quelque chose de précis dans une phrase, par exemple, *I want that one*;
- RB (Adverbe) : Avec 56 occurrences, est utilisé pour renforcer ou modifier un autre élément de la phrase, comme dans *It's not that difficult*;

- CJT (Conjonction pour les verbes) : Avec 2 636 occurrences, elle est utilisée pour relier une proposition subordonnée à une proposition principale, souvent après certains verbes, par exemple, *I think that you are right*.

L'histogramme ci-dessous fournit une vue d'ensemble de cette répartition par différentes catégories :

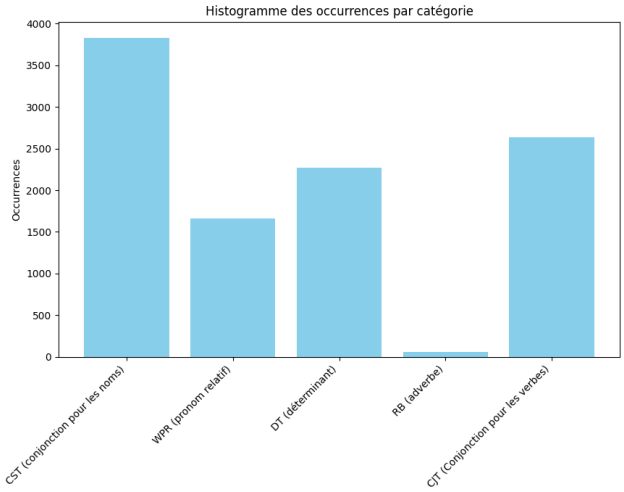

Figure 3: Visualisation de la répartition de *that* par catégorie

On peut voir que les catégories les plus fréquentes sont les conjonctions pour les noms (CST) et pour les verbes (CJT), suivies des déterminants (DT) et des pronoms relatifs (WPR), tandis que les adverbes (RB) sont nettement moins fréquents. Cela peut indiquer que les conjonctions et les déterminants jouent un rôle plus central dans la structure des phrases, tandis que les adverbes sont moins utilisés. A partir de ces résultats, on peut envisager par la suite d'ajuster la granularité du modèle d'étiquetage. Cette faible fréquence des adverbes pourrait également affecter l'étiquetage, en influençant la façon dont TreeTagger identifie les différentes fonctions de *that*.

Avant d'utiliser le corpus Brown pour entraîner TreeTagger, il est important de s'assurer qu'il a été prétraité correctement. Cela inclut les étapes suivantes :

- La tokenisation, c'est-à-dire diviser le texte en mots et en ponctuation;
- L'étiquetage initial des catégories grammaticales avec un modèle de base, comme le modèle Penn ou BNC;
- L'adaptation des étiquettes du corpus aux tags spécifiques du modèle choisi, comme le tagset C8 utilisé dans ce projet.

Il est crucial que les étiquettes du corpus soient compatibles avec celles du modèle afin d'assurer des performances optimales lors du réentraînement de TreeTagger.

4 Méthodologie

4.1 Évaluation des modèles pré-entraînés (BNC et Penn)

L'objectif est d'évaluer et de comparer la performance des modèles d'étiquetage morphosyntaxique BNC et Penn dans leur capacité à identifier et étiqueter correctement le mot *that* dans différentes fonctions grammaticales en anglais. Les résultats obtenus serviront de référence pour les améliorations futures.

4.1.1 Performance globale des taggers

- BNC : Une accuracy de 48%, ce qui signifie que le tagger BNC arrive quand même à identifier le type grammatical de *that* dans 48% des cas.
- Penn : Une accuracy plus faible de 27.5%, indiquant que ce tagger a rencontré plus de difficultés à classer correctement les occurrences de *that*.

BNC semble donc globalement plus performant que Penn sur cet échantillon de test.

4.1.2 Performances par Catégorie

Comme correspondance en fonction des différents taggers on a `WDT` qui est un pronom relatif interrogatif dans Penn, `PNQ` un pronom interrogatif dans BNC et `IN` une conjonction dans Penn.

On a ci-dessous un tableau récapitulatif des métriques obtenues :

Adverbe. Pour cette catégorie, on observe un faible rappel des deux modèles. Le modèle BNC affiche une précision et un rappel de 0 %, indiquant une totale incapacité à identifier *that* comme un adverbe. En revanche, le modèle Penn, malgré une précision maximale (100 %), présente un faible rappel de 5 %, suggérant qu'il identifie trop peu d'occurrences. Cette faiblesse pourrait être liée au fait que le modèle BNC n'a pas appris à reconnaître *that* comme un adverbe, tandis que le modèle Penn parvient à le détecter dans certains contextes, mais

Table 1: Résultats pour BNC

Label	Précision	Rappel	F1-score
RB	0.00	0.0	0.00
CJT	40.49	100.0	57.64
PNQ	0.00	0.0	0.00
DT0	60.13	92.0	72.73

Table 2: Résultats pour Penn

Label	Précision	Rappel	F1-score
WDT	73.08	19.0	30.16
IN	0.00	0.0	0.00
DT	97.73	86.0	91.49
RB	100.00	5.0	9.52

avec des résultats limités. Un affinement du modèle Penn ou l'exploration d'un autre modèle pourrait améliorer ces performances.

Conjonction. Le modèle BNC a un excellent rappel de 100 % mais une plus faible précision de 40,49 %. Un tel rappel signifie qu'il a bien identifié les cas où *that* est utilisé comme conjonction. Le modèle Penn, en revanche, n'a pas du tout pu identifier *that* comme une conjonction, avec un rappel et précision nuls. Cela suggère que Penn a une grande difficulté avec cette catégorie, peut-être aussi parce qu'il n'a pas appris à reconnaître *that* comme une conjonction. BNC semble particulièrement très bien adapté à cette catégorie, ce qui indique que son entraînement ou son jeu d'étiquettes est mieux aligné avec l'utilisation de *that* comme conjonction. Pour cette catégorie, on peut dire que le modèle BNC est meilleur.

Déterminant. Les deux modèles affichent des rappels élevés, avec BNC surpassant légèrement Penn. Ces résultats montrent que *that* est bien étiqueté comme déterminant dans la plupart des cas pour les deux modèles, bien qu'il y ait encore quelques petites erreurs. BNC présente une bonne performance avec 92% de rappel et 60,13 % de précision, tandis que Penn obtient une très bonne précision de 97,73 % et un rappel de 86 %, avec un F1-score de 91,49 %. La différence entre les deux modèles (92 % contre 86 %) est relativement faible, mais elle suggère que BNC gère légèrement mieux cette catégorie. Les erreurs peuvent être dues à des ambiguïtés dans le contexte ou à des cas où *that* pourrait être mal interprété comme un pronom ou une conjonction. Les deux modèles sont efficaces pour la catégorisation de *that* en tant que déterminant, avec BNC légèrement plus précis.

Pronom. Le modèle BNC montre une totale incapacité à identifier *that* comme pronom, ce qui montre une faiblesse significative pour cette catégorie avec un score de 0 % en précision et en rappel. Bien que le rappel de Penn soit relativement faible (19 %), il a tout de même réussi à identifier *that* comme pronom dans certains contextes, bien qu'avec un succès limité. Cette différence entre les deux taggers pourrait être due à un manque d'exemples de pronoms relatifs dans le corpus d'entraînement de BNC, tandis que Penn, bien qu'inferieur, semble mieux gérer ces cas. Toutefois, les deux modèles échouent largement à identifier *that* comme pronom, et un ajustement ou l'exploration d'autres modèles pourrait être nécessaire pour améliorer la performance.

4.1.3 Conclusion générale

Les résultats montrent que les modèles BNC et Penn sont tous deux efficaces dans certaines catégories, mais présentent des faiblesses importantes dans d'autres. BNC, avec une précision globale plus élevée (48 % contre 27,5 % pour Penn) est particulièrement performant dans la catégorie conjonction, mais échoue complètement dans les catégories adverbe et pronom. En revanche, Penn affiche de bonnes performances sur la catégorie déterminant, mais il se révèle totalement inefficace pour la catégorie conjonction. Ces résultats soulignent les lacunes des modèles dans les catégories adverbe et pronom interrogatif, où les deux modèles échouent presque totalement. BNC semble trop biaisé vers la catégorie conjonction, tandis que Penn manque de généralisation. Pour améliorer la classification de *that* on peut penser à affiner les modèles actuels, ou d'explorer de nouveaux modèles d'étiquetage, afin de mieux traiter les catégories où les modèles échouent, comme les adverbes et les pronoms.

4.1.4 Recommandations

- Pour les tâches où *that* est utilisé comme conjonction ou déterminant, le modèle BNC semble être le plus adapté.

- Pour la reconnaissance des pronoms, le modèle Penn offre une solution plus efficace, bien que les résultats demeurent faibles.

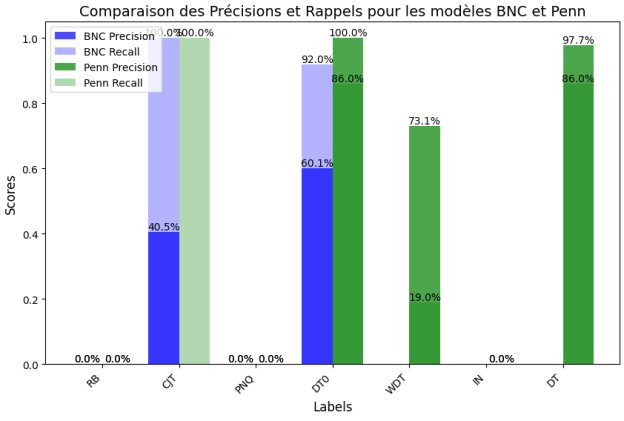

Figure 4: Comparaison des métriques des Modèles Penn et BNC

L'option **`-st .`** précise que le point (`.`) est utilisé comme marqueur de fin de phrase dans le corpus, ce qui est essentiel pour segmenter correctement les phrases lors de l'entraînement.

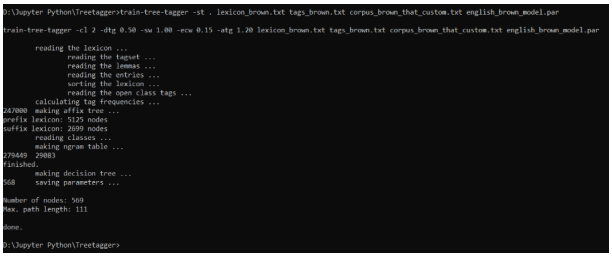

Figure 5: Réentraînement de TreeTagger

4.2 Réentraînement de TreeTagger

Dans cette partie, nous détaillons le processus de réentraînement de TreeTagger en utilisant les données extraites du corpus Brown.

L'entraînement d'un modèle avec **TreeTagger** repose sur la génération d'un modèle paramétrique à partir d'un corpus annoté. Ce processus est réalisé grâce à la commande suivante : `train-tree-tagger -st . lexicon_brown.txt tags_brown.txt corpus_brown_that_custom.txt english_brown_model.par`

Cette commande prend en entrée trois éléments essentiels :

- **Un lexique (`lexicon_brown.txt`)**, qui contient une liste de mots avec leurs catégories grammaticales possibles et, dans certains cas, leurs lemmes. Ce fichier permet d'améliorer la précision de l'étiquetage en associant directement les mots connus à leurs tags corrects.

- **Un ensemble d'étiquettes grammaticales (`tags_brown.txt`)**, qui répertorie tous les tags présents dans le corpus. Ce fichier garantit la cohérence de l'annotation en définissant les classes grammaticales reconnues par le modèle.

- **`corpus_brown_that_custom.txt`**, un corpus annoté, qui sert de base d'apprentissage. Il contient des phrases où chaque mot est associé à son tag grammatical, ce qui permet au modèle d'apprendre les structures et les relations entre les mots.

En sortie, la commande génère un fichier `english_brown_model.par`, qui constitue le modèle paramétrique entraîné. Ce modèle peut ensuite être utilisé pour l'analyse automatique de nouveaux textes, permettant d'étiqueter chaque mot en fonction de sa catégorie grammaticale. Il est conçu pour améliorer la précision de l'étiquetage, en particulier ici pour les distinctions grammaticales du mot **that**.

4.3 Fine - tuning

Plusieurs entraînements ont été effectués pour observer le changement dans les performances de chaque modèle.

L'objectif initial était d'ajuster le modèle en utilisant d'abord l'ensemble du corpus Brown, puis de réduire progressivement la taille des ensembles en fonction des catégories pour montrer à quel point la taille des données impacte sur la qualité de l'évaluation.

4.3.1 Premier Réentrainement

Le processus a démarré sur l'ensemble complet des données de Brown, avec un total d'occurrences pour chaque catégorie, permettant de tester pleinement les capacités du modèle.

4.3.2 Deuxième Réentrainement (1ère réduction des données)

L'étape suivante a vu une diminution des occurrences des différentes catégories. On a eu à prendre 52 occurrences d'adverbe, 3 249 occurrences de conjonction pour les noms, 1 506 occurrences de pronom relatif, 2 000 occurrences de conjonction pour les verbes, 1 500 occurrences de déterminant.

4.3.3 Troisième Réentrainement (2ième réduction des données)

L'ensemble de données a été encore diminué, offrant ainsi au modèle un plus petit nombre d'exemples pour chaque catégorie a l'exception de l'adverbe. On a donc pris RB 52 occurrences d'adverbe, 1 000 occurrences de conjonction pour les verbes, 1 500 occurrences de conjonction pour les noms, 2 000 occurrences de déterminant, 1 000 occurrences de pronom relatif.

4.3.4 Quatrième Réentrainement (3ème réduction progressive des données)

Dans cette étape, le jeu de données a été de nouveau diminué a l'exception de RB. Ceci avec, 52 occurrences d'adverbe, 500 occurrences de conjonction pour les verbes, 750 occurrences de conjonction pour les noms, 800 occurrences de déterminant, 500 occurrences de pronom relatif.

4.3.5 Cinquième Réentrainement (sous-ensemble restreint)

Enfin, on a terminé avec un plus petit sous-ensemble réduit des données. Cela permet de mieux comprendre comment chaque catégorie influe sur l'entraînement, sans surcharger le modèle. On a eu, entre autres, 52 occurrences d'adverbe, 100 occurrences de conjonction pour les verbes, 150 occurrences de conjonction pour les noms, 300 occurrences de déterminant et 350 occurrences de pronom relatif.

4.4 Évaluation et Discussion

Après avoir ajusté et optimisé le modèle avec différents sous-ensembles de données, il est important de l'évaluer. Cela permet de mesurer la précision de l'étiquetage et de vérifier comment le modèle fonctionne avec les différentes catégories et sous-ensembles. Cette étape aide à identifier les points forts et faibles du modèle, pour pouvoir l'améliorer.

Dans cette section, nous analyserons les performances du modèle avec différentes métriques et présenterons les résultats de l'évaluation selon les configurations de données

4.4.1 Évaluation - Réentrainement 1

Catégorie	Précision	Rappel	F1-score
RB	0.00%	0.00%	0.00%
CJT	68.57%	96.00%	80.00%
CST	39.84%	100.00%	56.98%
DT	96.39%	80.00%	87.43%
WPR	73.08%	19.00%	30.16%

Catégorie	Accuracy	Nb that	Bien tagués
RB	59.00%	56	0
CJT	59.00%	2636	96
CST	59.00%	3831	100
DT	59.00%	2272	80
WPR	59.00%	1662	19
Total	59.00%	10457	295

Table 3: Résultats de la première itération de l'entraînement

Adverbe. On peut voir que Le modèle ne parvient pas du tout à identifier les *that* adverbiaux, avec une précision, un rappel et un F1-score de 0%. Aucun des 100 cas de test n'a été correctement identifié, ce qui montre que le modèle n'a pas réussi à apprendre cette catégorie rare. Cette catégorie est peu représentée dans le corpus Brown, avec seulement 0,54% du total. La rareté de ces exemples rend l'apprentissage difficile, et il était probable que le modèle classe ces cas dans des catégories plus fréquentes. Ce déséquilibre entre la faible présence dans Brown (56 occurrences) et le set de test (100 cas) aggrave le problème.

Conjonction pour les verbes. Le modèle est performant dans l'identification des conjonctions, avec une précision de 68.57%, un rappel de 96% et un F1-score de 80%. Sur 100 vrais cas de CJT dans le test, 96 ont été correctement identifiés. Cependant, la précision plus faible indique la présence de faux positifs, où des instances non pertinentes sont classées comme des conjonctions. Cette catégorie, qui représente 25.21% du corpus Brown, est bien apprise grâce à un grand nombre d'exemples, mais le modèle pourrait être amélioré en équilibrant mieux les catégories pour éviter la sur-classification.

Conjonction pour les noms. On a un rappel parfait (100%) pour CST, identifiant tous ses vrais cas, mais une faible précision de 39.84%. Cela montre une sur-classification de cette catégorie, qui représente 36.64% du corpus

Brown. Le modèle semble souvent classer des cas incertains (fort bien) comme CST, ce qui explique les nombreux faux positifs. Même si le rappel est parfait, la faible précision montre que la sur-représentation de cette catégorie dans Brown a causé des erreurs.

Déterminant. Le modèle montre une bonne performance avec une précision de 96.39%, un rappel de 80% et un F1-score de 87.43%. Il identifie correctement 80 cas de DT sur 100, avec très peu de faux positifs. Représentant 21.73% du corpus Brown, cette catégorie est bien différenciée des autres, ce qui permet au modèle de bien apprendre et de maintenir un bon équilibre entre précision et rappel.

Pronom. Le modèle a du mal avec les pronoms, obtenant seulement un rappel de 19% et une précision de 73.08%. Cela montre qu'il manque beaucoup de pronoms dans les prédictions, avec peu de faux positifs. Sur 100 pronoms dans le test, seulement 19 ont été correctement identifiés. La faible représentation de cette catégorie dans le corpus (15.89%) et la confusion possible avec d'autres catégories comme CJT ou CST rendent cette tâche difficile pour le modèle.

Performance globale. On peut voir que le modèle arrive à identifier 295 catégories de **that** sur 500 phrases de test. L'accuracy globale de 59% montre que, bien que certaines catégories soient bien identifiées, des ajustements sont nécessaires. Le modèle est meilleur pour les catégories fréquentes (CJT, CST, DT) mais a du mal avec les rares (RB, WPR). Les déséquilibres dans le corpus, avec des catégories comme CST surreprésentées, causent des sur-classifications, tandis que les rares ne sont pas assez apprises. Une analyse des erreurs pourrait aider à comprendre les confusions. Pour améliorer la performance, il serait utile de rééquilibrer le corpus ou d'explorer d'autres modèles.

Après avoir analysé les métriques de performance, il est important d'examiner la matrice de confusion pour mieux comprendre les erreurs de classification du modèle.

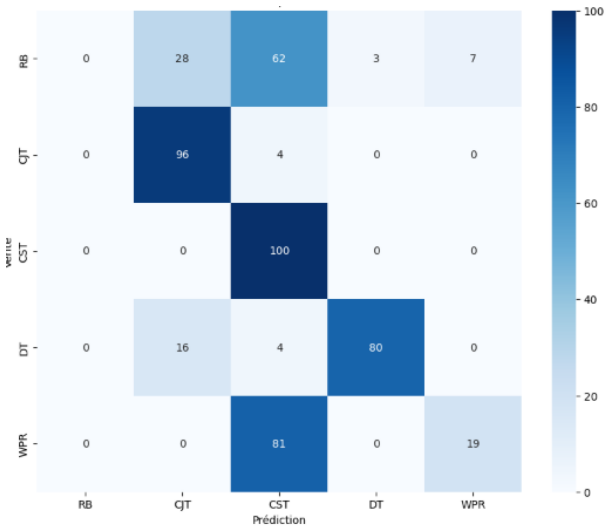

Figure 6: Matrice de Confusion 1

Adverbe RB. On peut voir que, aucun adverbe (RB) n'a été correctement identifié, sur 100 vrais RB, 28 sont classés comme CJT, 62 comme CST, 3 comme DT et 7 comme WPR. La confusion la plus importante est avec CST, suivi de CJT.

Conjonction CJT. Pour les conjonctions (CJT), 96 sur 100 sont bien classées, mais il y a des confusions avec RB et DT (28 RB et 16 DT sont classés comme CJT).

Conjonction CST. Rappel parfait mais faible précision. Le modèle identifie correctement tous les CST, mais classifie à tort beaucoup d'autres catégories comme CST, en particulier RB (62) et WPR (81).

Déterminant DT. Excellente précision et bon rappel. Le modèle est très précis dans l'identification des DT, avec peu de faux positifs. Cependant, il manque 20% des vrais DT.Le déterminant (DT) est bien identifié, mais 3 cas sont mal classés comme RB.

Pronom WPR. Enfin, pour le pronom, seulement 19 sur 100 sont bien classés, avec des erreurs fréquentes comme CJT et CST.

Ces erreurs montrent que le modèle a du mal à distinguer certaines catégories, notamment les adverbes et pronoms mais identifie bien les conjonctions et les déterminants.

4.4.2 Évaluation - Réentrainement 2

Catégorie	Précision	Rappel	F1-score
RB	100.00%	1.00%	1.98%
CJT	58.18%	96.00%	72.45%
CST	40.65%	100.00%	57.80%
DT	98.28%	57.00%	72.15%
WPR	66.67%	20.00%	30.77%
Catégorie	**Accuracy**	**Nb that**	**Bien tagués**
RB	54.80%	52	1
CJT	54.80%	2000	96
CST	54.80%	3249	100
DT	54.80%	1500	57
WPR	54.80%	1506	20
Total	54.80%	8307	274

Table 4: Résultats de la deuxième itération de l'entraînement

Performance globale. 274 sur 500 exemples de test de catégories *that* sont bien tagués. L'analyse des performances du modèle met en évidence une diminution globale de l'accuracy, passant de 59.00 % à 54.80 %. Cette baisse traduit une légère dégradation de la performance générale, suggérant que les ajustements apportés n'ont pas permis d'améliorer l'efficacité du système de classification. L'un des facteurs majeurs expliquant cette tendance est la réduction du nombre d'exemples dans le corpus d'entraînement Brown, ce qui a eu des répercussions variées selon les catégories.

Adverbe et Pronom. Les catégories Adverbe (RB) et Pronom (WPR) illustrent particulièrement les difficultés rencontrées avec les classes rares. Bien que l'on observe une amélioration marginale du rappel pour ces catégories (RB passant de 0 % à 1 %, WPR de 19 % à 20 %), ces progrès restent insuffisants pour garantir une classification efficace. Le modèle semble toujours incapable d'apprendre correctement les caractéristiques de ces classes, ce qui souligne la nécessité d'adopter des stratégies spécifiques pour pallier le déséquilibre des données.

Déterminant. L'évolution des performances met en évidence des changements notables dans l'équilibre entre précision et rappel. Par exemple, la catégorie Déterminant (DT) voit sa précision augmenter de 96.39 % à 98.28 %, mais cette amélioration se fait au détriment du rappel, qui chute brutalement de 80 % à 57 %.

Conjonction CJT. De même, pour la catégorie Conjonction (CJT), la précision baisse de 68.57 % à 58.18 %, tandis que le rappel reste stable à 96 %. Ces évolutions montrent que le modèle a ajusté sa stratégie de classification, parfois en privilégiant la précision au détriment du rappel, et inversement.

Conjonction CST. L'impact de la diminution du corpus d'entraînement se manifeste de manière hétérogène. Alors que certaines catégories, comme CST (Conjonction de subordination), maintiennent un rappel parfait de 100 % malgré une réduction de corpus (de 3831 à 3249 exemples), d'autres souffrent d'une forte dégradation des performances.

En particulier, la baisse du nombre d'exemples pour DT (de 2272 à 1500) et CJT (de 2636 à 2000) a considérablement affecté leur reconnaissance. Cela confirme que la taille et la qualité du corpus d'entraînement jouent un rôle déterminant dans la stabilité du modèle.

Enfin, les résultats mettent en lumière des confusions récurrentes entre certaines catégories, notamment entre CJT, CST et WPR. Cette tendance suggère que les frontières entre ces classes restent floues pour le modèle. En conséquence, il est nécessaire de développer des caractéristiques plus discriminantes pour renforcer la différenciation entre ces catégories et réduire les erreurs de classification.

Globalement, Ces résultats soulignent l'importance d'un corpus suffisant pour stabiliser le modèle.

Ci-dessous la matrice de confusion obtenue :

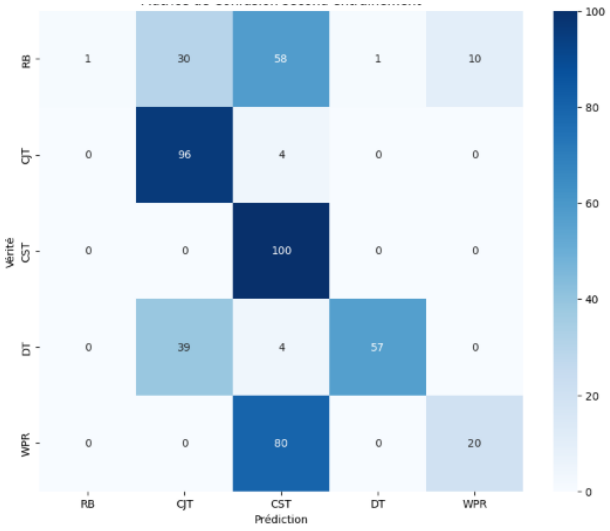

Figure 7: Matrice de Confusion 2

Adverbe. Le modèle parvient maintenant à identifier correctement un cas d'adverbe, ce qui est une légère amélioration par rapport à la première évaluation où aucun n'était identifié. Cependant, la performance reste très faible. Sur 100 vrais RB, 58 sont classés comme CST (contre 62

précédemment), 30 comme CJT (contre 28), 10 comme WPR (contre 7) et 1 comme DT (contre 3). La confusion principale reste avec CST, suivie de CJT.

Conjonction CJT. Le rappel reste excellent et identique à la première évaluation. Cependant, la précision a significativement diminué (de 68.57% à 58.18%). Le modèle continue à identifier correctement la plupart des CJT, mais la tendance à sur-classifier d'autres catégories comme CJT s'est accentuée, notamment pour RB et DT.

Conjonction CST. Le rappel parfait est maintenu, comme dans la première évaluation. La précision reste faible mais a légèrement augmenté (de 39.84% à 40.65%). Le modèle continue à classifier correctement tous les CST, mais classifie toujours à tort beaucoup d'autres catégories comme CST, en particulier RB et WPR.

Déterminant DT. La précision a augmenté (de 96.39% à 98.28%), mais au prix d'une baisse significative du rappel (de 80% à 57%). Le modèle est devenu plus conservateur dans ses prédictions de DT, faisant moins d'erreurs mais manquant beaucoup plus de vrais DT.

Pronom WPR. On observe une légère amélioration du rappel (de 19% à 20%) mais une baisse de la précision (de 73.08% à 66.67%). Le modèle identifie correctement un peu plus de WPR, mais fait plus d'erreurs dans cette identification. La confusion principale reste avec CST (80 cas mal classés, presque identique aux 81 cas de la première évaluation).

En conclusion, le modèle est bon pour les CST, mais reste mauvais pour les RB, WPR et DT, nécessitant un rééquilibrage et un ajustement global.

4.4.3 Évaluation - Réentrainement 3

Catégorie	Précision	Rappel	F1-score
RB	100.00%	1.00%	1.98%
CJT	73.08%	95.00%	82.61%
CST	40.00%	100.00%	57.14%
DT	90.91%	80.00%	85.11%
WPR	58.06%	18.00%	27.48%

Table 5: Résultats de la troisième itération de l'entraînement

Catégorie	Accuracy	Nb that	Bien tagués
RB	58.80%	52	1
CJT	58.80%	1000	95
CST	58.80%	1500	100
DT	58.80%	2000	80
WPR	58.80%	1000	18
Total	58.80%	5552	294

Table 6: Résultats de la troisième itération de l'entraînement

Performance globale. On observe une légère amélioration de l'accuracy, passant de 54.80% à 58.80%, rapprochant ainsi les résultats de la performance initiale de 59.00%. Bien que l'accuracy générale ait progressé, la réduction du corpus d'entraînement a eu des effets variés selon les catégories.

Adverbe RB. Les performances de la catégorie Adverbe restent très faibles malgré la réduction du corpus. Aucune évolution significative n'a été observée entre la 2e et la 3e évaluation, avec un rappel constamment bas (1.00%) et une précision parfaite basée sur un seul cas. Cette précision parfaite masque une réalité trompeuse, car elle repose sur une identification correcte d'un seul exemple. Cette catégorie continue d'être problématique, avec un faible rappel, et la réduction du corpus n'a pas amélioré la situation.

Conjonction CJT. Pour la catégorie Conjonction (CJT), la réduction du corpus a conduit à une amélioration des performances. La précision et le F1-score connaissent une nette progression dans la 3e évaluation, la précision passant de 58.18% à 73.08%, bien que le rappel diminue légèrement de 96.00% à 95.00%. La diminution du corpus semble paradoxalement avoir permis au modèle de mieux équilibrer la précision et le rappel, avec une performance améliorée. Le modèle a trouvé un compromis plus stable, suggérant une meilleure gestion des données d'entraînement.

Conjonction CST. La catégorie CST maintient un rappel parfait à 100% à chaque itération, mais la précision reste faible. Les performances restent stables à travers les trois évaluations, avec une légère fluctuation de la précision. La sur-classification des exemples CST demeure un problème, le modèle classant de nombreux cas incorrects comme CST, malgré la réduction du corpus.

Déterminant DT. La catégorie Déterminant (DT) a retrouvé un bon équilibre précision-rappel dans la 3e évaluation. La précision reste élevée

à 90.91%, et le rappel a récupéré à 80%, avec une amélioration notable par rapport à la 2e évaluation. L'augmentation du nombre d'exemples après la réduction semble avoir été bénéfique pour cette catégorie, permettant au modèle de mieux équilibrer la précision et le rappel.

Pronom WPR. Les résultats de la catégorie Pronom (WPR) montrent une tendance à la baisse, avec des performances en déclin sur les trois évaluations. La précision diminue continuellement, et le rappel reste faible, atteignant seulement 18% dans la 3e évaluation. La réduction du corpus a aggravé les performances de cette catégorie, qui reste un défi majeur pour le modèle.

Il est important d'explorer des stratégies pour améliorer les performances des catégories RB et WPR, comme un meilleur échantillonnage. Il faudrait aussi analyser la sur-classification de CST. Continuer à ajuster la distribution des exemples dans le corpus d'entraînement serait utile, en profitant des améliorations de CJT et DT. Une analyse détaillée des erreurs pourrait aider à comprendre les faiblesses du modèle et à corriger les problèmes de classification, d'où :

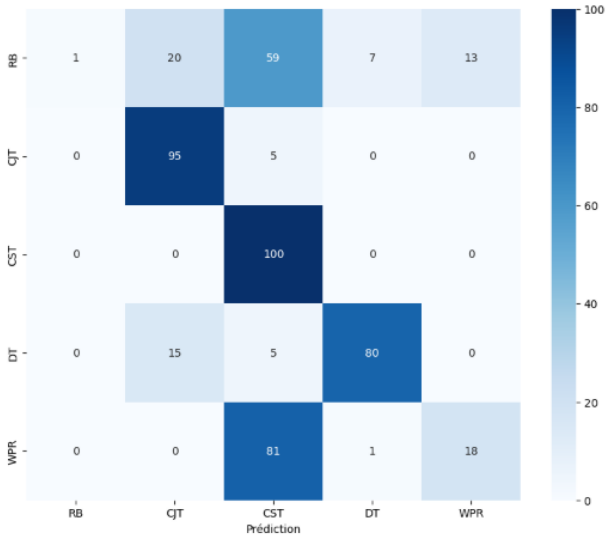

Figure 8: Matrice de Confusion 3

Adverbe RB. Le modèle identifie correctement un seul cas d'adverbe, comme dans la deuxième évaluation. La majorité des RB sont mal classifiés, 59 comme CST, 20 comme CJT, 13 comme WPR, et 7 comme DT. La confusion avec CST reste prédominante, mais on observe une augmentation de la confusion avec CJT et WPR par rapport aux évaluations précédentes.

Conjonction CST. Avec CST, on peut noter une persistance du problème de sur-classification, avec un rappel parfait mais une faible précision. La confusion principale reste avec WPR (81 cas) et RB (59 cas). On a une légère augmentation de la confusion avec CJT et DT par rapport à la 2e évaluation.

Conjonction CJT. Pour les conjonctions, la performance globale est améliorée avec un meilleur équilibre précision-rappel. On peut observer une réduction des faux positifs, notamment pour la catégorie DT (15 cas contre 39 dans la deuxième évaluation). On note une légère augmentation de la confusion avec CST (5 cas).

Déterminant DT. Les déterminants ont montré un meilleur équilibre précision-rappel par rapport à la deuxième évaluation. On note une réduction de la confusion avec CJT (15 cas contre 39 dans la 2e évaluation) et une légère augmentation de la confusion avec CST (5 cas) voir une apparition d'une confusion avec WPR (1 cas).

Pronom WPR. les pronoms ont une précision relativement faible (58,06%), avec une persistance d'une forte confusion avec CST (81 cas), une légère augmentation de la confusion avec RB (13 cas contre 10 dans la 2e évaluation) et une apparition d'une faible confusion avec DT (1 cas).

En résumé, le modèle est performant pour les CJT et DT, mais il reste peu fiable pour les RB et WPR.

4.4.4 Évaluation - Réentrainement 4

Catégorie	Précision	Rappel	F1-score
RB	100.00%	2.00%	3.92%
CJT	75.81%	94.00%	83.93%
CST	44.64%	100.00%	61.73%
DT	79.17%	76.00%	77.55%
WPR	61.11%	33.00%	42.86%

Catégorie	Accuracy	Nb that	Bien tagués
RB	61.00%	52	2
CJT	61.00%	500	94
CST	61.00%	750	100
DT	61.00%	800	76
WPR	61.00%	500	33
Total	61.00%	2602	305

Table 7: Résultats de la quatrième itération de l'entraînement

Adverbe RB. Lors de la quatrième évaluation, l'adverbe a montré une légère augmentation du

rappel, passant de 1% à 2%, tout en maintenant une précision de 100%. Cependant, ce chiffre reste trompeur, car il repose sur un nombre très faible d'exemples corrects (seulement 2 adverbes identifiés). Le modèle reste donc largement insuffisant, et bien que l'on constate un léger progrès, la faible représentation des adverbes dans le corpus de Brown limite l'apprentissage, ce qui explique cette performance encore très modeste.

Conjonction CJT. Pour la catégorie des conjonctions, une amélioration de la précision a été observée au fil des itérations, atteignant 75.81% à la quatrième évaluation, malgré une légère baisse du rappel (94%). Le modèle semble avoir trouvé un bon équilibre entre précision et rappel, avec un F1-score qui atteint son maximum (83.93%). La réduction du nombre d'exemples dans le corpus de Brown ne semble pas avoir affecté négativement les performances, suggérant qu'une sur-représentation des conjonctions dans le jeu de données initial pouvait avoir eu un impact sur l'apprentissage. Cela indique une bonne généralisation du modèle, qui identifie désormais plus précisément les conjonctions tout en maintenant un haut niveau de rappel.

Conjonction CST. Elle a montré une nette amélioration, avec une précision qui est passée de 40% à 44.64%, ce qui a permis d'augmenter le F1-score de 57.14% à 61.73%. Le rappel, qui reste à 100%, continue d'indiquer une sur-classification, mais l'amélioration de la précision suggère que le modèle parvient mieux à éviter les erreurs dans l'identification des conjonctions CST. Bien que ce problème de sur-classification persiste, sa réduction progressive témoigne de l'efficacité de la réduction de la taille du corpus d'entraînement.

Déterminants. En revanche, la catégorie des déterminants montre une dégradation de la performance, avec une baisse de la précision (de 90.91% à 79.17%) et du rappel (de 80% à 76%). Le F1-score a également diminué, ce qui suggère que le modèle a davantage de difficultés à identifier correctement les déterminants, produisant plus de faux positifs et de faux négatifs. La réduction du corpus pourrait être une explication de cette dégradation.

Pronom WPR. le modèle a montré une nette amélioration, avec un rappel passant de 18% à 33%, accompagné d'une hausse du F1-score de 27.48% à 42.86%. Bien que la précision n'ait augmenté que légèrement, l'amélioration du rappel indique que le modèle est désormais capable d'identifier une plus grande proportion de pronoms correctement, et ce, malgré la réduction du nombre d'exemples. Cela suggère un meilleur pouvoir de généralisation, un progrès notable pour une catégorie initialement difficile à traiter.

Les résultats indiquent que le modèle est plus performant sur certaines catégories que sur d'autres, ce qui met en lumière les défis persistants liés à la représentation inégale des classes dans le corpus de référence.

Ci - dessous la matrice de confusion résultante :

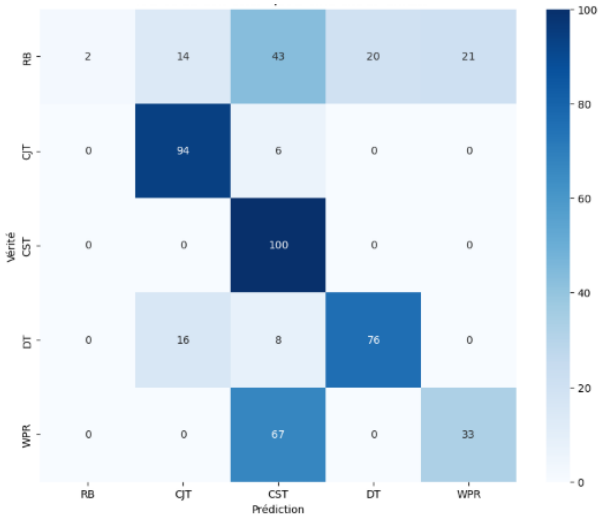

Figure 9: Matrice de Confusion 4

Adverbe RB. Cette matrice de confusion nous montre que, le modèle identifie correctement deux cas d'adverbes, contre un seul dans les deuxième et troisième évaluations. La confusion avec CST a diminué (43 cas contre 59 dans la 3e évaluation) ainsi que la confusion avec WPR et DT a augmenté (21 et 20 cas respectivement, contre 13 et 7 dans la 3e évaluation).

Conjonction CJT. Le nombre de CJT correctement identifiés a légèrement diminué (94 contre 95 dans la 3e évaluation). La confusion avec DT reste stable (16 cas contre 15 dans la 3e évaluation) et une nouvelle confusion significative avec RB (14 cas, non observée précédemment) est survenue.

Conjonction CST. La sur-classification persiste mais diminue légèrement (224 prédictions CST contre 250 dans la 3e évaluation). La confusion avec WPR a diminué (67 cas contre 81 dans la 3e évaluation) et la confusion avec RB a également diminué (43 cas contre 59 dans la 3e évaluation).

Déterminant. On a une baisse du nombre de DT correctement identifiés (76 contre 80 dans la 3e évaluation). On note aussi une augmentation de la confusion avec CJT (16 cas contre 15 dans la 3e

évaluation) et une nouvelle confusion importante avec RB (20 cas, non observée précédemment).

Pronom WPR. Il y a eu une augmentation significative des WPR correctement identifiés (33 contre 18 dans la 3e évaluation). Une réduction importante de la confusion avec CST (67 cas contre 81 dans la 3e évaluation) est notée ainsi qu'une nouvelle confusion avec RB (21 cas, non observée précédemment).

Ces changements dans les patterns de confusion suggèrent des ajustements du modèle qui ont apporté des améliorations, mais aussi de nouvelles difficultés, notamment dans la distinction des RB.

4.4.5 Évaluation - Réentrainement 5

Catégorie	Précision	Rappel	F1-score
RB	0.00%	0.00%	0.00%
CJT	74.36%	87.00%	80.18%
CST	43.86%	100.00%	60.98%
DT	79.59%	78.00%	78.79%
WPR	59.65%	34.00%	43.31%

Catégorie	Accuracy	Nb that	Bien tagués
RB	59.80%	52	0
CJT	59.80%	100	87
CST	59.80%	150	100
DT	59.80%	300	78
WPR	59.80%	350	34
Total	59.80%	500	299

Table 8: Résultats de la cinquième itération de l'entraînement

Adverbe RB. Dans la cinquième évaluation, les résultats pour les adverbes sont revenus à zéro, comme dans la première évaluation. Le modèle ne parvient plus à identifier les adverbes, contrairement aux précédentes évaluations où des progrès avaient été réalisés. Le nombre d'exemples dans le corpus est resté le même (52), mais la perte de performance montre que le modèle a du mal à apprendre correctement cette catégorie.

Conjonction CJT. Pour ces conjonctions, la précision a légèrement diminué, et le rappel a aussi baissé. Cependant, le modèle continue de bien performer, même avec une réduction importante des exemples dans le corpus (de 500 à 100). La baisse du rappel indique que le modèle commence à rater certaines conjonctions à cause de la réduction des données.

Conjonction CST. Pour les conjonctions CST, la précision a baissé un peu, mais le modèle continue de prédire presque toutes les conjonctions (rappel de 100%). Cela montre que le modèle fait encore des erreurs de classification, mais il continue de sur-classifier cette catégorie, c'est-à-dire qu'il classe plus de mots que nécessaire en CST. Ce problème de sur-classification est toujours là, même si le corpus a été réduit de manière importante.

Déterminant DT. Les résultats pour les déterminants sont restés assez stables, même avec moins d'exemples (de 800 à 300). La précision et le rappel n'ont pas changé de manière significative, ce qui montre que le modèle a appris à identifier les déterminants et qu'il est capable de le faire avec moins d'exemples.

Pronom WPR. Pour les pronoms, la précision a légèrement baissé par rapport à la quatrième évaluation, mais le rappel s'est légèrement amélioré passant de 33% à 34%. On note que le F1-score s'est amélioré, atteignant son meilleur niveau sur les cinq évaluations. Cela montre que le modèle est en train de mieux identifier les pronoms, probablement grâce à un meilleur équilibre dans le corpus d'exemples. Même avec moins de données, le modèle semble mieux s'adapter.

Conclusion. Le problème pour les adverbes est encore très marqué et doit être corrigé. Les autres catégories (conjonctions, déterminants, pronoms) ont montré des performances stables ou légèrement améliorées. Cependant, la sur-classification des conjonctions CST reste un problème à résoudre. Ces résultats soulignent l'importance de bien équilibrer les données d'entraînement pour chaque catégorie pour obtenir les meilleures performances.

On a obtenu la matrice de confusion suivante :

Adverbe RB. Aucun RB correctement identifié, contre 2 dans la 4e évaluation. Les 100 RB sont mal classifiés, 44 comme CST, 20 comme WPR, 19 comme CJT, et 17 comme DT. Le modèle a complètement perdu sa capacité à identifier les adverbes. La confusion avec CST reste prédominante, mais on observe une augmentation de la confusion avec CJT par rapport à la 4e évaluation.

Conjonction CJT. On a eu 87 vrais CJT correctement identifiés, contre 94 dans la 4e évaluation, 10 faux négatifs classés comme CST, 3 comme DT et 30 faux positifs 19 comme RB, 11

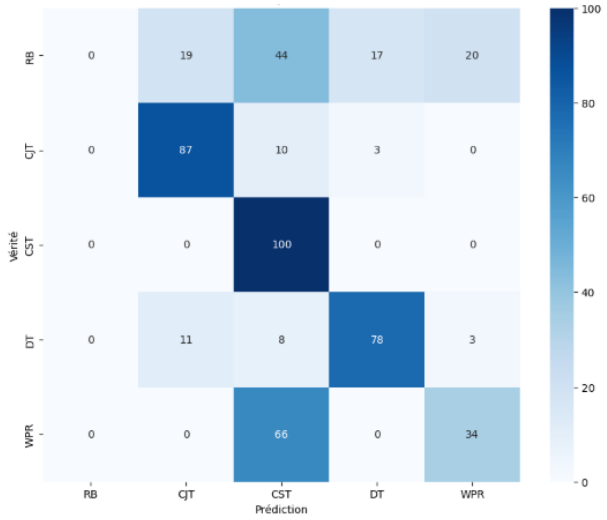

Figure 10: Matrice de Confusion 5

comme DT. Le modèle maintient une bonne performance pour CJT malgré la réduction du corpus. La confusion avec CST a légèrement augmenté, suggérant des difficultés à distinguer certaines conjonctions des CST. La réduction des faux positifs RB (19 contre 14 dans la quatrième évaluation) indique une meilleure distinction entre CJT et RB.

Conjonction CST. Tous les vrais CST (100) sont correctement identifiés. On observe 128 faux positifs, 66 WPR, 44 RB, 10 CJT, 8 DT. Le problème de sur-classification persiste, mais avec une légère amélioration. La confusion principale reste avec WPR et RB, comme dans les évaluations précédentes. On a aussi une réduction des faux positifs RB (44 contre 43 dans la 4e évaluation) et WPR (66 contre 67). La réduction du corpus n'a pas résolu le problème de sur-classification des CST.

Déterminant DT. On a eu 78 vrais DT correctement identifiés, contre 76 dans la quatrième évaluation. 22 faux négatifs ont été observé, 11 classés comme CJT, 8 comme CST, 3 comme WPR ainsi que 20 faux positifs dont 17 RB, 3 CJT. On note Réduction de la confusion avec RB (17 faux positifs contre 20 dans la 4e évaluation). Le modèle montre une légère amélioration dans l'identification des déterminants. La confusion avec RB a diminué, suggérant une meilleure distinction entre ces deux catégories. L'apparition de faux négatifs classés comme WPR est nouvelle et mérite attention.

Pronom WPR. 34 vrais WPR sont correctement identifiés, contre 33 dans la 4e évaluation, 66 faux négatifs, tous classés comme CST ainsi que 23 faux positifs, dont 20 RB, 3 DT. On observe un maintien d'une forte confusion avec CST (66 faux négatifs, identique à la quatrième évaluation). Le modèle maintient sa performance améliorée pour WPR, observée depuis la quatrième évaluation. La confusion persistante avec CST reste un défi majeur. L'apparition de faux positifs DT est nouvelle et indique une difficulté croissante à distinguer certains déterminants des pronoms.

4.4.6 Conclusion - Évaluations des différents entraînements

Ordre	Réentraînement	Accuracy
1	4ème réentraînement	61%
2	5ème réentraînement	59.80%
3	1er réentraînement	59%
4	3ème réentraînement	58.80%
5	2ème réentraînement	54.80%

Réentraînement	Nb that
4ème réentraînement	2602
5ème réentraînement	952
1er réentraînement	10457
3ème réentraînement	5552
2ème réentraînement	8307

Table 9: Comparaison des performances des différents réentraînements

À l'issue des cinq réentraînements effectués, il apparaît que le quatrième entraînement se distingue par la meilleure performance, atteignant 61%, malgré un nombre relativement faible d'occurrences de *that* (2 602). En comparaison, bien que le cinquième entraînement présente une performance légèrement inférieure (59,80%), il repose sur un nombre d'occurrences beaucoup plus réduit (952). De même, le premier entraînement, avec 10 457 occurrences, affiche une performance de 59%, ce qui suggère que l'augmentation du nombre d'occurrences ne garantit pas nécessairement une meilleure performance. Par ailleurs, le troisième entraînement (58,80% avec 5 552 occurrences) et le deuxième (54,80% avec 8 307 occurrences) confirment cette tendance. Ainsi, il apparaît que la performance du modèle ne dépend pas uniquement du volume des occurrences de *that*, mais probablement d'autres facteurs influençant son efficacité.

4.5 Évaluation avec Stanza et UDPipe

Nous allons dans cette partie, utiliser les modèles prétrainés **spacy_stanza** et **spacy_udpipe** pour ef-

fectuer des évaluations sur nos données de test. Ces modèles sont accessibles via spaCy, une bibliothèque populaire de traitement du langage naturel.

SpaCy_UDPipe est un package qui permet d'intégrer le modèle UDPipe à spaCy. Il effectue des tâches de prétraitement linguistique comme l'étiquetage des parties du discours (POS tagging), l'analyse des dépendances syntaxiques et la lemmatisation. Ce modèle est multilingue, léger, et conforme aux standards Universal Dependencies (UD), bien qu'il puisse être moins précis pour certaines langues par rapport à d'autres solutions spécialisées.

SpaCy_Stanza, quant à lui, permet d'intégrer Stanza à spaCy. Stanza est une bibliothèque puissante développée par le Stanford NLP Group, qui offre des modèles basés sur des réseaux de neurones profonds pour diverses tâches avancées telles que le POS tagging, les dépendances syntaxiques, la lemmatisation, et la reconnaissance d'entités nommées (NER). Bien que plus précise dans certaines langues, Stanza peut être plus lente et nécessite plus de ressources comparé à **spacy_udpipe**.

Nous avons donc évalué les performances du modèle spaCy-UDPipe sur nos fichiers de test (au format .txt cités plus haut données de test) et avons obtenu les résultats suivants :

Catégorie	Précision	Nb test
RB	11.0	100
CJT	96.0	100
CST	83.0	100
DT	89.0	100
WPR	87.0	100
Global	73.2	500

Catégorie	Bien tagués
RB	11
CJT	96
CST	83
DT	89
WPR	87
Global	366

Table 10: Évaluation des résultats avec spaCy-UDPipe

Nous avons aussi évalué les performances du modèle spaCy_stanza sur les mêmes fichiers de test et avons obtenu les résultats suivant :

Catégorie	Précision	Nb test
RB	67.0	100
CJT	100.0	100
CST	83.0	100
DT	97.0	100
WPR	100.0	100
Global	89.4	500

Catégorie	Bien tagués
RB	67
CJT	100
CST	83
DT	97
WPR	100
Global	447

Table 11: Évaluation des résultats spaCy-Stanza

Une étude comparative a été réalisée avec les performances du quatrième réentrainement car il avait de meilleurs performances sur le corpus Brown.

Récapitulons les performances du quatrième réentrainement avec le corpus Brown :

Catégorie	Nb test	Bien pred Brown
RB	100	2
CJT	100	94
CST	100	100
DT	100	76
WPR	100	33
Global	500	305

Catégorie	Nb that Brown	Précision Brown
RB	52	2.0
CJT	500	94.0
CST	750	100.0
DT	800	76.0
WPR	500	33.0
Global	2602	61.0

Table 12: Récapitulatif des performances sur Brown

Toutes nos performances sont donc obtenues sur nos dits fichiers.txt de test. Les résultats obtenus sont les suivants :

Catégorie	Modèle	Précision	Rappel	F1-score	Accuracy	Nb that	Nb test	Nb correct
RB	PENN	100.00	5.0	9.52	27.5	-	100	5
RB	BNC	0.00	0.0	0.00	48.0	-	100	0
RB	Brown	100.00	2.0	3.92	61.0	52	100	2
RB	spaCy-UDPipe	100.00	11.0	19.82	11.0	-	100	11
RB	spaCy-Stanza	100.00	67.0	80.24	67.0	-	100	67
CJT	PENN	0.00	0.0	0.00	27.5	-	100	0
CJT	BNC	40.49	100.0	57.64	48.0	-	100	100
CJT	Brown	75.81	94.0	83.93	61.0	500	100	94
CJT	spaCy-UDPipe	100.00	96.0	97.96	96.0	-	100	96
CJT	spaCy-Stanza	100.00	100.0	100.00	100.0	-	100	100
CST	PENN	0.00	0.0	0.00	27.5	-	100	0
CST	BNC	40.49	100.0	57.64	48.0	-	100	100
CST	Brown	44.64	100.0	61.73	61.0	750	100	100
CST	spaCy-UDPipe	100.00	83.0	90.71	83.0	-	100	83
CST	spaCy-Stanza	100.00	83.0	90.71	83.0	-	100	83
DT	PENN	97.73	86.0	91.49	27.5	-	100	86
DT	BNC	60.13	92.0	72.73	48.0	-	100	92
DT	Brown	79.17	76.0	77.55	61.0	800	100	76
DT	spaCy-UDPipe	100.00	89.0	94.18	89.0	-	100	89
DT	spaCy-Stanza	100.00	97.0	98.48	97.0	-	100	97
WPR	PENN	73.08	19.0	30.16	27.5	-	100	19
WPR	BNC	0.00	0.0	0.00	48.0	-	100	0
WPR	Brown	61.11	33.0	42.86	61.0	500	100	33
WPR	spaCy-UDPipe	100.00	87.0	93.05	87.0	-	100	87
WPR	spaCy-Stanza	100.00	100.0	100.00	100.0	-	100	100

Table 13: Performance des modèles pour chaque catégorie

L'analyse détaillée des performances des modèles Brown, spaCy-UDPipe, spaCy-Stanza, Penn, BNC, Brown révèle des différences marquées selon les catégories grammaticales.

Adverbe RB. Pour les adverbes, SpaCy-Stanza surpasse nettement les autres modèles dans l'identification des adverbes, avec un rappel de 67% tout en maintenant une précision parfaite. Les autres modèles, bien qu'ayant une précision parfaite (sauf BNC), ont un rappel très faible, indiquant qu'ils manquent la majorité des adverbes. BNC échoue complètement dans cette catégorie. L'accuracy élevée de Brown (61%) malgré son faible F1-score suggère qu'il performe mieux sur d'autres catégories.

Conjonction CJT. Concernant les conjonctions (CJT), SpaCy-Stanza atteint une performance parfaite, suivi de près par spaCy-UDPipe. Brown montre une bonne performance, avec un bon équilibre entre précision et rappel. BNC a un rappel parfait mais une faible précision, suggérant une sur-classification. PENN échoue complètement dans cette catégorie.

Conjonction CST. En ce qui concerne les conjonctions CST, SpaCy-UDPipe et spaCy-Stanza montrent des performances identiques et supérieures. Brown et BNC ont un rappel parfait mais une faible précision, indiquant une tendance à sur-classifier cette catégorie. PENN échoue à nouveau complètement.

Déterminant DT. Pour les déterminants, tous les modèles performent relativement bien dans cette catégorie, avec spaCy-Stanza en tête. PENN montre une performance surprenamment bonne, surpassant même Brown. La différence entre les modèles est moins marquée que pour d'autres catégories.

Pronom WPR. Quant aux pronoms, SpaCy-Stanza atteint une performance parfaite, suivi de près par spaCy-UDPipe. Il y a un écart important avec les autres modèles. Brown et PENN montrent des performances moyennes, tandis que BNC échoue complètement.

Globalement, spaCy-Stanza excelle dans toutes les catégories et offre la meilleure performance globale, suivi de près par spaCy-UDPipe. Brown présente de bonnes performances dans des catégories spécifiques, notamment les conjonctions CST, mais échoue dans d'autres, comme les adverbes et les pronoms. PENN est performant sur les déterminants, mais échoue complètement sur certaines catégories. BNC Montre des faiblesses majeures, avec des échecs complets sur plusieurs catégories.

Ces résultats suggèrent que pour une analyse générale, spaCy-Stanza est le modèle optimal, tandis que pour des tâches spécifiques, le modèle Brown peut être privilégié, et spaCy-UDPipe constitue un bon compromis.

Contributions :

- Welehela Taweuteu Orchelle Patricia (Tech Lead) : Développeuse, Rédaction rapport
  - Réalisation des différents réentrainements de treetagger sur le corpus brown + analyse et comparaison des modèles entraînés sur divers splits du corpus d'entraînement et interprétation des résultats;
- Ngasseu Ndifo Lyse Priscille : Rédaction Rapport, Développeuse;
  - Description, test et comparaison des modèles BNC et Penn et interprétation des résultats;
  - Réannotarion du corpus brown;
- Kpatoukpa Kpodjro : Rédaction Rapport, Développeur
  - Description des modèles Spacy_UDpipe et Spacy_Stanza;
  - Test, comparaison des modèles Spacy_UDpipe et Spacy_Stanza et interprétation des résultats;
- Ndiaye bassirou Serigne : Rédaction Rapport, Développeur
  - Description, Visualisation, Analyse morphosyntaxique, et analyse lexicale du corpus de Brown;
  - Étude comparative des 05 modèles.

Références

- Sigogne, M. (2009). *Master 2 en Linguistique - Analyse automatique des structures syntaxiques et du marquage morphologique*. http://infolingu.univ-mlv.fr/english/Bibliographie/Articles/Sigogne_2009_Master2.pdf

- Sajous, F. (2020). *TreeTagger - Sans Interface*. http://fsajous.free.fr/SDL/SL0720X/treetagger/SL0720-sajous-TreeTagger_sansInterface.pdf
- Alignalco. (n.d.). *TreeTagger*. http://alignalco.free.fr/alignalco1/treetagger
- RNTI Editions. (n.d.). *RNTI - Publication scientifique*. https://editions-rnti.fr/render_pdf.php?p=1002281
- Hamon, P. (2010). *Syntaxe et Structure des Données Linguistiques*. https://hal.science/hal-00493847v1/document
- Hamon, P. (2014). *Cours M2 IBM-TAL - Syntaxe*. https://perso.limsi.fr/hamon/Teaching/P5/M2IBM-TAL-2014-2015/Cours/21-syntaxe.pdf
- Nouvels, D. (2017). *Statistiques Linguistiques - Présentation*. https://damien.nouvels.net/cours/statscorp/04_StatistiquesLinguistiques-pres.pdf
- ResearchGate. (2017). *Expérience d'entraînement de TreeTagger et d'intégration à l'interface Web de SATO*. https://www.researchgate.net/publication/333185349_Experience_d'entrainement_de_TreeTagger_et_d'integration_a_l'interface_Web_de_SATO
- Aiakide. (2022). *Corpus Annotation TD5*. https://pro.aiakide.net/cours/2022Annot/Corpus%20annotation%20-%20TD%205.pdf
- Zoumbara, R., Diwérsy, O., Ouedraogo, M., & Martin. (2020). *JADT 2020 - Text Mining*. https://agritrop.cirad.fr/597679/1/ZOUMBARA_ROCHE_DIWERSY_OUEDRAOGO_MARTIN_JADT2020.pdf
- Dumas, F. (2011). *Approche de la Linguistique Computationnelle*. https://dumas.ccsd.cnrs.fr/dumas-00631517/document
- Renater. (2020). *FAQ TreeTagger*. https://groupes.renater.fr/wiki/txm-users/public/faq
- DataCamp. (2020). *What is a Confusion Matrix in Machine Learning*. https://www.datacamp.com/fr/tutorial/what-is-a-confusion-matrix-in-machine-learning
- Innovatiana. (2020). *Understanding the Confusion Matrix in AI*. https://www.innovatiana.com/post/understand-confusion-matrix-in-ai
- Picsellia. (2020). *Matrice de Confusion en Computer Vision*. https://www.picsellia.fr/post/matrice-confusion-computer-vision
- Google. (2020). *Thresholding in Classification*. https://developers.google.com/machine-learning/crash-course/classification/thresholding?hl=fr
- LinkedIn. (2020). *How can you interpret classification models' confusion?*. https://fr.linkedin.com/advice/3/how-can-you-interpret-classification-models-confu lang=fr
- Microsoft. (2020). *Confusion Matrix and Classification Metrics*. https://learn.microsoft.com/fr-fr/dynamics365/finance/finance-insights/confusion-matrix
- Scikit-Learn. (2020). *Confusion Matrix and Metrics Evaluation*. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html