6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TreeTagger entraîné avec le *Critical Pronouncing Dictionary* de J. Walker face aux textes modernes

Francois Huang, Blanche Miret, Preethi Srinivasan, Dao Thauvin
RELIA Recherche En Licence Informatique et études Anglophones
Encadrants : Jean-Baptiste Yunès et Nicolas Ballier
Université de Paris, 5 rue Thomas Mann, 75013, Paris
blanche.miret@etu.univ-paris-diderot.fr,
francois.huang@etu.univ-paris-diderot.fr,preethi.lfp@gmail.com,
dao.thauvin@etu.univ-paris-diderot.fr

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

RÉSUMÉ

TreeTagger (Helmut Schmid) est un outil moderne d'annotation de texte, par des lemmes et des catégories grammaticales. L'objectif de cette recherche est de déterminer si cet outil est capable d'assimiler les catégories grammaticales utilisées au 18ème siècle. Pour ce faire, nous avons utilisé le *Critical Pronouncing Dictionary* de John Walker (1791) afin de récupérer des catégories grammaticales datant du 18ème siècle des différents mots présents dans la langue anglaise pour obtenir un tagset. Ensuite nous avons créé deux fichiers .par entraînés avec des phrases du *Brown Corpus*. L'un utilise le tagset obtenu précédemment et un lexique extrait du dictionnaire de John Walker et l'autre utilise un jeu d'étiquettes et un lexique extraits du *Brown Corpus*. À partir des fichiers .par, nous avons laissé notre outil analyser certains textes modernes provenant du *Brown Corpus* de la bibliothèque NLTK et une partie du dictionnaire de John Walker. Sur des mêmes fichiers test, nous aboutissons à une précision de 33.5% en moyenne (32% de précision sur un texte provenant du dictionnaire de Walker et 35% de précision sur un texte provenant du *Brown Corpus*) avec le fichier .par utilisant les tags présents dans le dictionnaire de John Walker alors que la précision avec le fichier .par créé à partir du *Brown Corpus* est de 93.5% en moyenne (91% de précision sur un texte provenant du dictionnaire de Walker et 96% de précision sur un texte provenant du *Brown Corpus*), ce qui nous amène à penser que les tags du 18eme siècle ne sont pas adaptés à l'annotation de texte avec TreeTagger. Cependant, l'entraînement de TreeTagger et les expériences ont été effectués sur une faible quantité de données, et notre méthode pour utiliser les tags du 18ème nécessite une traduction des tags du 18ème siècle en tags de Brown Corpus. Nous perdons donc certains tags spécifiques du dictionnaire de Walker. En améliorant ces aspects, les résultats peuvent différer. Une analyse qualitative a par ailleurs montré l'incohérence de certaines étiquettes de Walker.

MOTS-CLÉS : TreeTagger, Walker, catégorie grammaticale, 18ème siècle.

KEYWORDS: TreeTagger, Walker, Part-Of-Speech Tag, 18th century.

REMERCIEMENTS

Nous remercions en premier lieu Nicolas Ballier et Jean-Baptiste Yunès qui nous ont accompagnés tout au long de ce projet et introduits aux règles de la recherche académique. Nous remercions également Nicolas Trapateau pour son travail effectué sur le *Critical Pronouncing Dictionary* de John Walker que nous avons utilisé comme jeu de données dans nos recherches, ainsi que les organisateurs de la conférence des Apprenti-e-s Chercheur-euse-s 2020 de nous donner la possibilité de partager ce travail. Enfin, merci au programme IdEx de l'Université de Paris grâce à qui ce projet fut financé.

Références

FRANCIS W. N. & KUCERA H. (1979). *Brown corpus manual*. Letters to the Editor.

HUANG F., MIRET B., SRINIVASAN P. & THAUVIN D. (2020). Github repository. https://github.com/daothauvin/TreeTaggerWithWalker.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, p. 44–49.

TRAPATEAU N. (2015). *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIIIe siècle sur la base du témoignage des dictionnaires de prononciation, des vers et de la musique vocale*. Thèse de doctorat, Université de Poitiers.

WALKER J. (1791). *A Critical Pronouncing Dictionary*. British Library.