TreeTagger entrainé 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

RÉSUMÉ

TreeTagger 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 des phrases du 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 et ainsi entrainer TreeTagger. 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. Nous aboutissons à une précision de 34% en moyenne alors que la précision avec les tags présent dans le Brown Corpus est de 93%, ce qui nous amène à penser que TreeTagger n'est pas adapté à l'annotation de texte avec des tags du 18 siècle. Cependant, l'entrainement de TreeTagger et les expériences ont été effectué 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.

ABSTRACT

TreeTagger trained with *Critical Pronouncing Dictionary* by J. Walker against modern corpus

TreeTagger is a modern tool for text annotation, using lemmas and grammatical categories. The objective of this research is to determine whether this tool is capable of assimilating the grammatical categories of 18th century sentences. To do this, we used the *Critical Pronouncing Dictionary* by John Walker (1791) to retrieve 18th century grammatical categories of the different words present in the English language and thus train TreeTagger. We let our tool analyze some modern texts from the *Brown Corpus* of the NLTK library and part of John Walker's dictionary. We reach an accuracy of 34% on average while the accuracy with the tags present in the Brown Corpus is 93%, which leads us to think that TreeTagger is not adapted to annotate text with tags from the 18th century. However, TreeTagger training and experiments have been performed on a small amount of data, and our method to use 18th century tags requires a translation of 18th century tags into Brown Corpus tags. We therefore lose some specific tags from Walker's dictionary. As we improve these aspects, the results may differ.

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

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